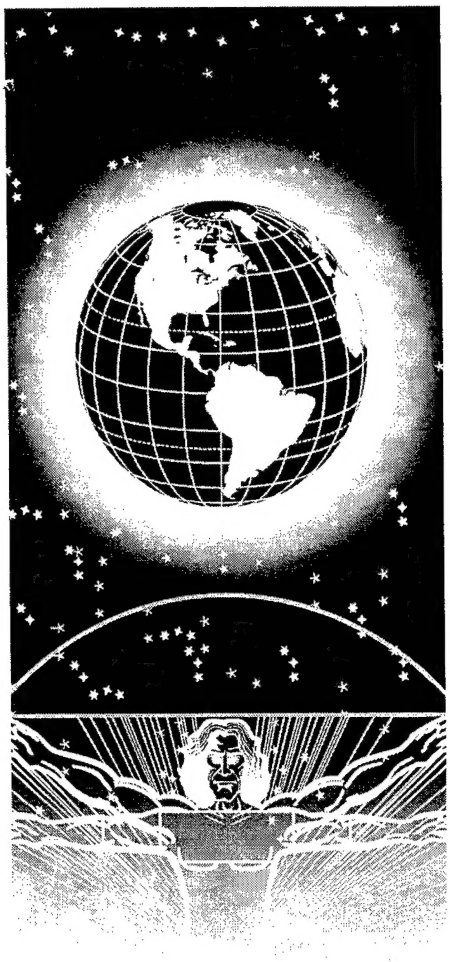
AFRL-HE-AZ-TR-2000-0126



UNITED STATES AIR FORCE RESEARCH LABORATORY

FACTORS THAT INFLUENCE THE EFFECTIVENESS OF TRAINING IN ORGANIZATIONS: A REVIEW AND META-ANALYSIS

Winston R. Bennett, Jr.

Air Force Research Laboratory
Warfighter Training Research Division
6030 South Kent Street
Mesa AZ 85212-6061

Winford Arthur, Jr.

Texas A&M University
Department of Psychology
College Station TX 77843-4235

JUNE 2001

AIR FORCE MATERIEL COMMAND
AIR FORCE RESEARCH LABORATORY
Human Effectiveness Directorate
Warfighter Training Research Division
6030 South Kent Street
Mesa AZ 85212-6061

NOTICES

Publication of this report does not constitute approval or disapproval of the ideas or findings. It is published in the interest of STINFO exchange.

Using Government drawings, specifications, or other data included in this document for any purpose other than Government-related procurement does not in any way obligate the US Government. The fact that the Government formulated or supplied the drawings, specifications, or other data, does not license the holder or any other person or corporation, or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

WINSTON R. BENNETT JR.
Project Scientist

DEE H. ANDREWS
Technical Advisor

JERALD L. STRAW, Colonel, USAF
Chief, Warfighter Training Research Division

Copies of this report may be requested from:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218

http://stinet.dtic.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
July 2001	Final	Sep 1993 to Dec 1995

4. TITLE AND SUBTITLE

Factors that Influence the Effectiveness of Training in Organizations: A Review and Meta-Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Winston R. Bennett Jr.
Winfred Arthur, Jr.

5d. PROJECT NUMBER
1123

5e. TASK NUMBER
A2

5f. WORK UNIT NUMBER
19

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Warfighter Training Research Div
6030 South Kent Street
Mesa AZ 85212-6061

Texas A&M University
Department of Psychology
College Station TX 77843-4235

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Warfighter Training Research Div
6030 South Kent Street
Mesa AZ 85212-6061

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-AZ-TR-2000-0126

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT We examined training literature focusing on potential factors that could influence effectiveness of training interventions, and used meta-analysis to quantify the impact of these factors. A total of 177 articles and 466 independent data points were used. Across all 466 data points, training was found to be more effective than expected. In addition there was sufficient evidence (SD - .540) to support a search for hypothesized moderators of training effectiveness. Results were expected to contribute to the resolution of issues associated with the impact of aptitude-treatment interactions and the existence of positive-findings bias associated with the methodological rigor or quality of the evaluation study. Implementation quality was found to be a significant moderator of training effectiveness. Although the majority of studies (93%) did not report any needs assessment activities, those that did were found to be markedly more effective than those that did not. Results related to criteria used to evaluate training were not as expected. Effect size for training did not systematically vary as a function of the "level" of criteria and observed effect size for results criteria was substantially larger than expected. Different training methods were effective for different skills and tasks and functioned as moderators of training effectiveness. Quantitative indicators of the relative overall effectiveness of a variety of training methods, the overall effectiveness of training for various skills and tasks to be trained, and relative effectiveness of the methods for the training specific skills and tasks were calculated. Finally, no empirical evidence supporting the existence of aptitude treatment interactions or the presence of positive-findings bias in studies of training effectiveness was

15. SUBJECT TERMS
Aptitude-Treatment Interactions; Meta-analysis; Methodological rigor; Task characteristics; Training criteria; Training effectiveness; Training efficiency; Training evaluations; Training methods

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr Winston Bennett Jr
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED	120	19b. TELEPHONE NUMBER (include area code) 480.988.6561 x-297 DSN 474-6297

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The work presented in this report was accomplished in partial fulfillment of the requirements for the first author's Doctoral degree at Texas A&M University, College Station, Texas. The work described in this report would not have been possible without the considerable support of a number of individuals and organizations.

The authors wish to thank Ms. Christina Rodriguez and Mr. Michael Kaminski for their willingness to help gather the identified studies for use in this effort. Also, Ms. Gloria Koenig provided significant technical editing of the final manuscript for this publication.

The first author would also like to recognize the following individuals for their continuous support and encouragement of the first author during the conduct of this study.

Dr. Dave Woehr and Dr. Robert Pritchard for their strong support and encouragement during and after my tenure at Texas A&M. In addition, Ms. Pamela Stanush served as the other meta-analysis coder. She also worked very hard to help ensure that the individual study coding was complete and accurate and kept the database updated despite frequent changes, additions, and deletions. This study was extremely demanding and her hard work was a key component in its completion.

The men and women of the Armstrong Laboratory Human Resources Directorate are also acknowledged. This report would not have been possible without their continuous support and encouragement. In particular, Lieutenant Colonel James Bushman, Technical Training Research Division Chief, was a continually strong advocate of the work at Texas A&M. Further, Dr. Nestor K. Ovalle, Lieutenant Colonel William E. Wimpee, and Major Archie M. Smith made considerable sacrifices, as immediate supervisors, to provide the time and resources necessary to complete this work. The encouragement and support for this effort provided by Dr. Mark Teachout, at various points during this study is also greatly appreciated. Finally, Colonel William Strickland, the Director of the Human Resources Directorate, our Chief Scientist, Dr. Scientist, Dr. William Alley, and Dr. R. Bruce Gould, Technical Director for the Technical Training Research Division, were continually supportive of this effort and strongly encouraged and facilitated its successful completion.

Last, but by no means least, Drs. Henk Ruck, Jimmy Mitchell, and the late Dave Vaughan, are each due tremendous thanks for their friendship, mentoring, advice, encouragement, and frequent phone and electronic mail conversations throughout this effort.

PREFACE

This report documents work accomplished in partial fulfillment of the requirements for a doctoral degree at Texas A&M University and performed at the former Armstrong Laboratory, Human Resources Directorate, Mission Critical Skills Division, at Brooks Air Force Base, TX, under Work Unit 1123-A2-19, Contributive Research in Training Technologies.

Publication of this report was delayed due to personnel reassignments and laboratory reorganization. It is being published by the Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Training Research Division, in Mesa AZ, in the interest of scientific and technical information exchange.

INTRODUCTION

The continual need for individual and organizational development can be traced to numerous demands. These include maintaining superiority in the marketplace, enhancing employee skill and knowledge, and increasing productivity (Cascio, 1991). One of the most pervasive methods for enhancing the productivity of individuals and communicating organizational goals to new personnel is training. Training can be defined as a learning experience which is planned by the organization. This learning experience occurs after the trainee is a member of the organization and is designed to advance organizational goals (Campbell, Dunnette, Lawler, & Weick, 1970). Additionally, training is usually undertaken to produce a "relatively permanent" change in knowledge, skills, behavior, and attitudes (Cascio, 1991, p. 361).

At present, public and private sector organizations spend over 40 billion dollars annually on training-related activities (Eurich, 1985; Huber, 1985). Given the potential impact of training upon organizations and the costs associated with development and implementation of training, it is prudent to explore issues related to the efficacy of various training and development techniques and identify areas for future research and theoretical advancement.

During the past 12 years, there have been several reviews of the training and development literature (e.g., Goldstein, 1980; Latham, 1988; Tannenbaum & Yukl, 1992; Wexley, 1984). The period since 1980 represents a highly active period in the scientific training literature which has been marked by improvements in training design and development, and improvements in the empirical assessment of the effectiveness of training. There has also been a continued identification of critical issues affecting the effectiveness of training interventions.

The purpose of this report is to examine the training literature focusing on factors that could potentially influence the effectiveness of training interventions. These factors are namely: (a) the quality of the implementation of the training program, (b) different criterion measures of training effectiveness, (c) the use of different training methods, (d) the match between training methods and skill and task characteristics, (e) trainee characteristics, and (f) the methodological rigor/empirical design of training effectiveness studies.

The literature examined in this study was limited to the published scientific literature, including technical reports and conference reports/presentations, in training and development in organizational contexts from 1960 to 1993. Relevant practitioner-oriented applications of training method and theory, although not a primary focus of this report, have been cited and discussed where applicable. However, similar to past training and development reviews (e.g., Latham, 1988; Tannenbaum & Yukl, 1991; Wexley, 1984), the practitioner-oriented literature was excluded unless it met the inclusion criteria outlined later in this report. In most cases,

however, the practitioner-oriented literature was excluded because it tends to be driven by technological fads and does not advance theory or take advantage of scientific research outcomes in training development and evaluation (Goldstein, 1980; Latham, 1988).

The objectives of the present report are, therefore, to (1) summarize the literature related to each of the previously specified factors; (2) illustrate why each factor is important in terms of potentially influencing the effectiveness of training interventions; and (3) demonstrate how each factor influences training effectiveness.

REVIEW OF THE LITERATURE

Current Issues in Training Research

Over the past twenty years there have been five cumulative reviews of the training and development literature (Campbell, 1971; Goldstein, 1980; Latham, 1988; Tannenbaum & Yukl, 1992; Wexley, 1984). In the first review (Campbell, 1971), the training literature was characterized as (a) being driven by current fads; (b) atheoretical in approach; and (c) lacking empirical substance in evaluation methods. Campbell (1971) further stated that future research needed to evaluate training in terms of the differential effects of different strategies and to evaluate training outcomes in organizational terms.

Almost 10 years after Campbell's review, Goldstein (1980) concluded that training and development research was still somewhat atheoretical and nonempirical in nature. He did note, however, that there was some movement toward a more empirical focus in the development and evaluation of training. Highlighted research needs included a desperate need to evaluate the usefulness of different training techniques and a need to assess the impact of training upon employee attitudes, learning, behavior, and organizational goals (Goldstein, 1980).

The third review of the training literature (Wexley, 1984) cited over 150 recent empirical studies evaluating training and development activities. In addition to a continued emphasis on more rigorous empirical evaluations, several new challenges to training and development were highlighted. These included a need to address the effects of changing technology and cultural differences on training approaches and the adaptation of training methods for fundamental skills training in a growing population of non-english speaking workers (Wexley, 1984). This review highlighted the need for continued research in needs assessment at the organizational, task, and individual level and the need to evaluate the efficacy of training methods at these levels in pre-, post-, and then-paradigms (Wexley, 1984).

The fourth review, Latham (1988), emphasized a continued increase in empirical research studies published since the third review. It was also noted that many of the earlier concerns related to training theory had been realized. Latham (1988) argued that the continuing difficulty evidenced in the literature since the last review was "the inability of training research to bring about relatively permanent changes in the behavior of the practitioner" (p. 546). The fourth review also discussed training research conducted outside the United States and training which was focused on specific content such as leadership. Finally, Latham (1984) proposed the addition of a fourth category, demographic analysis, to the traditional needs assessment trichotomy of organizational, task, and person analyses (McGehee & Thayer, 1961). Demographic analysis would provide information related to training needs for populations of workers. Examples of studies emphasizing a demographic focus included perceptions of workers

over 40 years of age (Tucker, 1985); male and female managers (Berryman-Fink, 1985); and economically disadvantaged women (Streker-Seeborg, Seeborg, & Zegeye, 1984).

The fifth review (Tannenbaum & Yukl, 1992), provided the latest cumulation of the training and development literature. There were a number of critical issues which were highlighted in this latest comprehensive review. Training needs assessment continued to be an important component of training development in organizations. In fact, Saari, Johnson, McLaughlin, and Zimmerle (1988) demonstrated that only 27% of the companies they surveyed had any type of procedures for identifying training objectives and needs. In addition, Tannenbaum and Yukl's (1992) review highlighted advances in specific training methods such as simulations and games, high-technology methods, and behavior modeling. The continued need to develop and use more rigorous training evaluation methods and problems associated with the measurement of change due to training, were also highlighted and discussed.

The emergence of training issues in pre- and post-training environments and the continued evolution of training for special populations such as managers and teams were also discussed. With respect to pre- and post-training environments, the authors emphasized the importance of environmental cues and signals, trainee choice in the training situation, and the characteristics of the post-training environment for successful training outcomes. Finally, Tannenbaum and Yukl (1992) called for a "paradigm shift" to research designed to assess "why, when, and for whom a particular type of training is effective" (p. 433). The present study attempted to address issues associated with this paradigm shift.

The purpose of the previous discussion was to highlight key research issues that have been identified by training researchers. While a number of issues were identified, their impact, based on results from previous primary research studies, has not been adequately addressed. The present report attempts to use the extent primary research literature to determine the impact of these issues on training effectiveness. Also, given the potential impact of the issues discussed in the previous overview, the present report examined the training effectiveness literature from 1960 to 1993 focussing on the following factors that could potentially influence the effectiveness of training interventions: (a) the quality of the implementation of the training program, (b) different criterion measures of training effectiveness, (c) the use of different training methods, (d) the match between training methods and skill and task characteristics, (e) trainee characteristics, and (f) the methodological rigor/empirical design of training effectiveness studies. These factors were examined in terms of why each is important to the evaluation of training and how each factor could potentially influence the effectiveness of training interventions. The present study attempted to use the available scientific literature to assess the impact of these factors on the effectiveness of training. Thus, the present study also attempted, to some extent, to use the

existing literature to assess when, and for whom, a particular type of training is effective (Tannenbaum & Yukl, 1992). Each of the identified factors are discussed, in detail, in later sections of this report.

The following steps were used to accomplish the research objectives of this study. First, a qualitative review of the literature related to each of these factors was conducted. Second, a conceptual framework of the relationship of these factors to training effectiveness was developed. Third, several hypotheses related to the nature of the impact of these factors as potentially influencing training effectiveness were developed and tested using meta-analytic techniques. The use of meta-analytic techniques permitted a quantitative assessment of the influence of each proposed factor.

FACTORS THAT INFLUENCE TRAINING EFFECTIVENESS IN ORGANIZATIONS: A REVIEW AND META-ANALYSIS

The previous discussion of past training literature reviews highlighted a number of key factors that are common throughout much of the training research literature. In the following sections, each factor is described in detail, its importance discussed, and its role in terms of how it could potentially influence training effectiveness outcomes is highlighted.

Quality of the Implementation of Training Programs

Central to any intervention activity is the quality of the implementation of that intervention. "Quality of intervention" refers to the steps which are taken during the identification, development, execution, and evaluation of the intervention. As such, these initial steps are crucial for effective training programs. For example, Noe and Schmitt (1986) demonstrated that trainees who reacted positively to a skill needs assessment procedure were more likely to react favorably to the training program than their counterparts who disagreed with the assessment of their skill needs prior to training. The manner in which training and intervention programs are implemented has been a focus of research on the aspects of "quality" in implementation (Connell, Turner, & Mason, 1985; Pentz, Trebnow, Hansen, MacKinnon, Dwyer, Johnson, Flay, Daniels, & Cormack, 1990). Several definitions of implementation quality have been proposed. In the clinical program intervention literature, where intervention programs are targeted at high risk groups for such problems as substance abuse, implementation quality has been defined in one of three ways: (a) adherence, which refers to the fact that the intervention program is given to the experimental group and not to the control group; (b) exposure (or program fidelity), which refers to the quantity of the program delivered to a target group of individuals; and (c) reinvention, which is the extent to which the implemented program deviates from a pre-established course of action (Pentz, et al., 1990).

It would appear to be important to consider a number of issues when implementing an intervention program. These include assessing the feasibility of developing and implementing a program in an organization (e.g., a clinical or schoolhouse setting), identifying the content and activities to be focused upon in the intervention program, and identifying the needs of individuals who will participate in the intervention program (see Connell et al., 1985; Pentz et al., 1990).

In many ways, the implementation issues highlighted in the clinical intervention program literature are very similar to those found in the implementation of organizational training and development programs. Quality of implementation in a training program can be seen as related to the traditional trichotomy of systematic training needs assessment: organizational, task, and person analysis (McGehee & Thayer, 1961). According to Wexley and Latham (1991), these

three aspects of a traditional needs assessment seem to address specific questions related to training development, implementation, and evaluation. These questions are the following: What are the training needs in the organization and where is training needed? What are the requirements of the job that need to be learned in training? Who needs training and what kind of training is required? The extent to which each of these questions is addressed in the development of a training program can be used as an indicator of the quality of implementation.

Organization analysis refers to the examination of the organization as a whole, before any training development occurs. The rationale for an organization analysis is to identify what the training needs of the organization are and where the training is needed (Goldstein, 1993). A second goal of this level of analysis is related to identifying both short-term and more long-term organizational goals which may be addressed by a training program (Wexley & Latham, 1991). Further, organization analysis helps to determine the extent to which resources, in terms of people to be trained and support for their training, will be available. Finally, organization analysis can be used to assess the climate of the organization. Assessing the climate of the organization will help to identify the congruence between employee needs and goals, and those of the organization. This will be especially important in determining the success or failure of a training program to provide tangible benefits (e.g., changes in job behavior and organizational results) (Baldwin & Ford, 1988; Goldstein, 1993).

The second aspect of a systematic needs assessment is task analysis. Task analysis refers to the analysis of jobs to determine the tasks that are involved in performing the job that will be the focus of training. In addition, task analysis identifies the knowledge, skills, and abilities (KSAs) that are required to perform the tasks in the jobs to be trained. The identification of the KSAs leads to the specification of course objectives, behavioral objectives to be used in criterion development, and ultimately to the design of the training program (e.g., what training methods to use, course length, number of trainers, and number of trainees) (Wexley & Latham, 1991).

A third and final aspect of a systematic needs assessment is related to person analysis. Person analysis focuses upon the individuals who will be trained. In order to determine who needs the training and how much training they need, this analysis examines how the individual is currently performing their job. This can be determined from performance appraisals, behavioral observations of the person in their job, or through the use of proficiency tests of job knowledge. In addition, self-reports offer another way to assess the need for training. Ford and Noe (1987) developed and used a Need-For-Training Questionnaire for managers and supervisors in an organization. The respondents were required to review a list of skills required in their jobs and to rate themselves on the extent to which they felt they needed training on those skills. Thus, self-

reports, in combination with performance appraisal information, can be used to identify individuals for training (Goldstein, 1993).

Conducting a systematic needs assessment can heavily impact the overall quality of the implementation of a training program. Therefore, the extent to which a training intervention uses a systematic approach to needs assessment can influence the overall effectiveness of training. A systematic needs assessment can be used to specify a number of key features for the implementation (input) and evaluation (outcomes) of a training program. According to Baldwin and Ford (1988) "[training] research has concentrated on input factors that might affect training transfer rather than focusing on the appropriate measurement of the conditions of transfer" (p. 94).

In summary, the quality of the implementation process can be critical to the success of the training program. This is due to the fact that a systematic needs assessment provides the mechanism whereby the questions essential to successful training programs can be answered. That is, assessing the training needs of the organization, identifying job requirements to be trained, identifying who needs training, and the kind of training to be delivered, should result in more effective training programs. Therefore, studies which report a comprehensive needs assessment (e.g., person, task, and organizational analysis) are seen as being of "higher" quality, in terms of their implementation. Consequently, it is expected that training programs that are of higher quality, in terms of their implementation, should report larger effect sizes than those of lower quality.

A second factor that could potentially influence the effectiveness of training is the criteria used for evaluating training outcomes. The development and use of criteria for evaluating training and the potential problems associated with current training criteria is examined in the following sections.

Criterion Development

In training evaluation, as in any organizational program assessment, the ultimate goal of the enterprise is to demonstrate that the use of the program provided beneficial outcomes or changes to the organization. Evidence of the beneficial outcomes of a particular program, is based upon the analysis of criteria. Criteria can be best described as standards which can be used to evaluate either individuals (as in personnel selection and performance appraisal) or organizational programs and interventions (e.g., training and other organizational development [OD] programs) (Cascio, 1992; James, 1973). Consequently, criteria are operational statements of organizational goals or outcomes (Cascio, 1991).

In the case of training evaluation, the goal of criterion development is twofold. First, criteria can be used as measures to assess the relationship between performance in a training

program and subsequent performance on-the-job. Second, criteria can be used to determine if one training program was more beneficial than another.

A number of issues are especially salient when developing appropriate criteria for evaluating training. These issues include the extent to which selected criteria are relevant, deficient, contaminated, and reliable. Each of these is examined in turn. Criteria are relevant when the measures of success in the training program are related to performance on the target task (Goldstein, 1993; Thorndike, 1949). The process of conducting a needs assessment identifies the critical tasks, or behavioral objectives, to be performed on the job. Because, the requirements of the job should also be a central focus of the training, the criteria chosen should represent both the critical tasks and/or the behavioral objectives of the job. Thus, the relationship between the training objectives and the chosen evaluation criteria is an indication of the relevance of the criteria (Cascio, 1991; Goldstein, 1993).

Criteria are deficient when some aspects of the objectives identified in the needs assessment as being important for job performance are not included in the criteria used in the evaluation of training. One important way to reduce the occurrence of criterion deficiency is through the use of multiple criteria. Multiple criteria ensure a broader coverage of the objectives required for performance on the job. In addition, multiple criteria allow the program to be evaluated at the individual, workgroup, and organizational level. Criteria associated with performance in each of these areas need to be specified and assessed.

Criterion contamination occurs when extraneous elements present in the evaluation criteria, cause it to be less representative of the constructs identified in the training needs assessment (Goldstein, 1993). These extraneous elements can cause the conclusions regarding the validity of a training program to be incorrect. Criterion contamination can result from several sources of bias: opportunity bias, group characteristic bias, and knowledge of training performance (Cascio, 1991; Goldstein, 1993).

For example, opportunity bias occurs when a selected group of individuals has received training on new techniques for jet engine maintenance. When these individuals return to the workplace, they are given work assignments on the more complex maintenance activities, ostensibly because they are perceived to be "more capable" as a result of the training. In this example, training would be judged to be more valid since these workers are given more complex assignments. That is, an assessment of behavioral criteria related to the performance of the trained tasks would be artificially enhanced by the increased opportunities for the trainees to perform more complex tasks.

Another potential source of criterion contamination is also related to opportunities to perform trained tasks. It is called group-characteristic bias (Goldstein, 1993). Group-

characteristic bias occurs when a trained group does not have opportunities to perform the trained task, due to social or policy limitations in the workplace. Continuing the jet-engine maintenance example, suppose a group of trainees has recently completed a training course on the actual maintenance of jet engines. These individuals have spent a considerable amount of time learning about jet engines and how to maintain them. However, when the newly trained individuals arrive at their jobs they are not "allowed" to perform the tasks they have learned, because they are "new". In this example, the lack of opportunities to perform the trained task will reduce the effectiveness of the training as measured by job behavior criteria. In this instance, it would be important to assess the nature of the post-training environment and develop a set of criteria that are related to that environment so that these dynamics would be known and alternate measures of job performance could be developed (e.g., job knowledge tests, climate surveys) and used.

A final source of criterion contamination is related to the expectations of the training evaluators. This is known as knowledge of training performance (Goldstein, 1993). Knowledge of training performance occurs when the individual(s) evaluating the training program outcomes are familiar with the goals and objectives of the training and, therefore, selectively observe behaviors that are consistent with those of the training program. Two ways to avoid this source of contamination involve the use of multiple measures of work performance (e.g., performance checks, performance appraisals, and job knowledge tests) and a combination of subjective and objective measures obtained from independent sources.

A further issue in developing criteria for the evaluation of training is related to the consistency or reliability of the measures over time. Appropriate criteria should be stable across time. That is, ratings of performance using the same criteria should be stable at different points in time. However, it should be noted that criteria that are stable must also be relevant. If the selected criteria have no relevance to the measurement of outcomes due to training then their reliability is of little consequence.

When selecting and using criteria for evaluation, several other issues should also be addressed. These issues are related to the dimensionality of criteria. Criteria can possess temporal, static, and/or dynamic dimensionality (Ghiselli, 1956). We will examine each of these aspects of criteria.

Temporal dimensionality refers to the fact that criterion measures may not be independent of time. That is, criterion measurements taken at different points in time may produce strikingly different outcomes. To adequately address problems associated with temporal dimensionality, criteria should be identified for measurement at different points in time. Two special cases of temporal dimensionality are static and dynamic dimensionality. Each of these will be discussed e next.

Static dimensionality refers to the fact that criteria which are often used in evaluation are those which are measured at one point in time (Cascio, 1991). The assumption is that measuring a single criterion at one point in time adequately assesses and describes the employee's performance. For example, training evaluators measure an individual's mastery of trained material using an end-of-course test and assume that the score on such a test is indicative of performance in the future. Similarly, once trainees return to their workplace, they are given a performance check related to the content that was the focus of the training. Again, the assumption here is that trainees' performance on the performance check has adequately captured and described their performance. Cascio (1991) points out, this may not be the case. What is needed are criteria that sample job behaviors at multiple points in time to obtain a representative picture of employee performance and change.

Dynamic dimensionality is related to the fact that criteria may be dynamic, and of changing importance over time (Cascio, 1991). That is, criterion measures of performance taken early in an employee's job tenure, may not be related to job performance in the future. As the employee gains experience on the job, the dimensions of their performance that were relevant early in their job may not be as relevant for later job performance. In a study exploring several characteristics of dynamic criteria, Barrett, Caldwell, and Alexander (1985) stated that dynamic criteria may be manifested in any one of three possible forms: (a) changes in group average performance over time, (b) changes in validity over time, and (c) changes in the rank-ordering of performance scores on the criterion over time. Barrett et al., (1985) found little substantial evidence for significant change in criterion validities over time and that rank-ordering of scores over time appeared to be quite stable. However, the impact of criterion dimensionality is still an issue of considerable debate.

The summary conclusions of Barrett et al., (1985) have been challenged by Austin, Humphreys, and Hulin (1989). Austin et al., (1989) contend that the most appropriate approach to understanding criteria is to conduct research on criteria as criteria as opposed to examining artifacts as they suggest Barrett et al., (1985) had done. Also, Hulin, Henry, and Noon (1990) reviewed a number of studies and concluded that predictive validity was unstable and decreased over time. In a subsequent reanalysis of Hulin et al.'s, (1990) reported data, Barrett, Alexander, and Doverspike (1992) demonstrated that the conclusions drawn by Hulin et al., (1990) might be premature. The reanalysis conducted by Barrett et al., (1992) suggested that the problems of dynamic criteria and changing validities over time are worthy of exploration, but the results of Hulin et al., (1990) were not particularly relevant to predictive validities in a real world context of job performance.

In conclusion, Barrett et al., (1992) recommend that researchers need to be more rigorous in developing predictors and evaluation criteria and to focus on design studies to evaluate predictive validity over time. However, the question of the stability (or instability) of predictive validity over time remains unanswered.

As previously mentioned, the ultimate goal of the criteria is to demonstrate a benefit to the organization from the program. However, in the short term this may not be possible (Cascio, 1991). What is needed are criteria that are identified on the basis of the outcomes from a needs assessment. Further, selected criteria need to be of several types: (a) criteria that are immediate or proximal (e.g., an end-of-training course grade); (b) criteria that are intermediate (e.g., a four to six month performance check, work sample tests, or peer evaluations); and ultimately, (c) criteria that are summary or more long-term aggregates of aspects of the other criteria (Cascio, 1991). In the training evaluation literature, these can be viewed as being synonymous to learning criteria, behavioral criteria, and results criteria (see Goldstein, 1993).

In summary, there are numerous issues which should be considered in developing appropriate criteria for job performance *and* for training evaluation. In many cases, a training program may be seen as unsuccessful simply because inappropriate criteria were chosen as measures of training outcomes. Given the preceding discussion of criterion issues, it seems prudent to examine training criteria to determine their appropriateness as measures of training outcomes.

Outcome measures of training effectiveness can be seen as potentially influencing the observed effectiveness of training. In many cases, the most easily accessible outcome measures of performance may not be the most appropriate measures of training effectiveness. Moreover, it is reasonable to ask whether different criteria will provide different information. In other words, we may see differential outcomes from training as a function of the criteria chosen to measure effectiveness. For example, as the distance between the training event and criteria used to measure effectiveness increases, so does the likelihood that training benefits will not be observed. This is due to the fact that with increasing distance between the intervention and the measurement of outcomes, there are many variables that can intervene and reduce the observed effectiveness of training. These variables can include such things as social or organizational support, the availability of resources, and/or opportunities to perform trained tasks. The potential impact of criterion issues upon the assessment of training effectiveness are discussed in the following section.

Measures of Training Effectiveness. Kirkpatrick (1959; 1987) identified four broad types or categories of measures of training effectiveness. These have been labeled: reaction,

learning, behavior, and results. Historically, Kirkpatrick's topology has received widespread acceptance within the industrial/organizational (I/O) psychology community (Cascio, 1991). Although much of the published literature has referred to these four types as levels, the use of "levels" inaccurately implies a hierarchical relationship among the levels. The relationship among measures across types of criteria can be seen as a function, not of the other levels, but of the objectives of the training itself. Studies examining the dependence among the types of criteria (e.g., Alliger & Janak, 1989: Noe & Schmitt, 1986) have found limited empirical support for the existence of dependence. Additionally, these criteria types may be viewed from the organization's perspective as proceeding from proximal (e.g., individual or workgroup productivity) to more distal (e.g., absenteeism, turnover, utility metrics) outcomes from the training activity. Each of the measures in Kirkpatrick's topology is discussed in detail.

Reaction Measures. Reaction measures are concerned with trainees' feelings or impressions of the training. Such concerns as how they enjoyed the training experience, the approach taken in the training, and the accommodations provided during the training activity are typical of reaction measures. These measures are usually easy to obtain and are fairly nonintrusive. Furthermore, reaction measures can be used to assess trainee's perceptions of the usefulness of the training to the job (Cascio, 1991) or to their personal and professional growth. These impressions of trainees are usually not empirically evaluated in a pre- and post-training design. It is important to note that *liking* a training program is not the same thing as *learning* the content of the training. Furthermore, positive reactions to training may, in fact, not be related to learning and/or subsequent changes in behaviors and job performance improvement (Kaplan & Pascoe, 1977). In cases where training is designed as a reward for good performance or as a means of increasing employee pride (Alliger & Janak, 1989; Goldstein, 1993), reactions can be considered as an important and sufficient evaluative criteria. If personnel are happy or feel better about the organization at the end of training then it was "successful".

Alliger and Janak (1989) found almost no relationship between trainee reactions and the other levels in the topology, although their conclusions were based on a small number of studies. More recently, Mathieu, Tannenbaum, and Salas (1992) demonstrated that trainee reactions actually moderated the relationship between motivation and learning. In other words, trainees who reacted positively to the training were more motivated to learn from the training program. In their review of criterion issues in training evaluation, Tannenbaum and Yukl (1992) conclude "at this point, however, reaction measures are not a suitable surrogate for other indexes of training effectiveness" (p. 425).

In summary, reaction measures are the most proximal criteria to use for evaluating training outcomes. As such, reaction measures are minimally impacted by such things as the organizational environment, resource availability, and group-characteristic bias noted in the earlier discussion of criteria development. They may, nevertheless be deficient as criteria, since they typically do not reflect components of performance that are related to the job. However, because of their proximal nature, most training studies will use trainee reactions as criteria of effectiveness. Furthermore, it is expected that training programs which use reaction measures as criteria, will obtain larger effect sizes than those that use other, more distal criteria such as learning, behaviors, and results.

Learning Measures. Learning measures are objective, quantifiable measures of the learning outcomes of the training. Learning measures can take the form of formal examinations, performance checks, and peer evaluations. These are not, however, measures of job performance. Learning measures can, and should be, developed from a needs assessment and should be obtained as part of an evaluation design incorporating some amount of control so that learning outcomes can be directly attributed to the training objectives, not necessarily to the reactions of trainees. In addition, learning may not be manifest in subsequent job behaviors. This does not mean that the training was not beneficial. More distal criteria, such as behaviors, are susceptible to environmental variables that can influence the use of trained skills or capabilities. The post-training environment may not provide opportunities for the learned material to be performed (Ford, Quinones, Sego, & Speer Sorra, 1992). In fact, changes in learning measures may occur without changes in behavior, although this relationship should be explored. According to Tannenbaum and Yukl (1992) "trainee learning appears to be a necessary but not sufficient prerequisite for behavior change" (p. 425).

In summary, learning measures are more proximal criteria than either behaviors or results criteria, as measures of training effectiveness. As such, learning measures are also less susceptible to organizational intervening variables such as the social environment of the workplace or supervisory support of trained tasks. Also, learning criteria are usually easy to obtain and fairly nonintrusive in terms of their impact on work activities since they are typically obtained at the end of the training activities. Thus, it is expected that training programs which use learning criteria as measures of effectiveness, will report larger effect sizes than those that use either behavior or results criteria.

An examination of criteria used for evaluating training effectiveness must consider the impact of intervening organizational variables and the potential impact of these variables upon more distal criteria such as behaviors or results. As stated earlier, the timing associated with the

measurement of the criteria, from the point of the intervention, can be impacted by numerous characteristics of the post-training environment. With respect to training evaluation, the specification of distal criteria related to behaviors, and further, to results, must include an assessment of the favorability of the environment for the transfer of trained skills (Ford et al., 1992; Noe, 1986; Thayer & Teachout, 1993).

Behavior Measures. Behavior measures are measures of actual on-the-job performance. Behavior measures can be used to identify the effects of training on actual work performance. However, learning and behavior are not necessarily related. Severin (1952) found that the relationship between training-related learning and production records was quite small ($r = .11$). End-of-training measures (learning) should not be assumed to reflect future on-the-job performance (behaviors) unless the relationship between them has been empirically established. As stated earlier, the post-training environment plays a key role in providing opportunities to perform new behaviors which are developed as an outcome of a training intervention. Ford et al., (1992) studied Air Force jet engine mechanics after the completion of training. In a longitudinal study of the effects of training, they found significant differences in the post-training environments of the trainees. There were significant differences in the opportunities to perform trained skills and in some cases, there were significant delays before the trainee first performed the trained tasks. The impact of trainee perceptions of the favorability of the post-training environment is discussed further in the next section.

Environmental favorability refers to trainees' perceptions of the work environment. These perceptions are related to characteristics of the workgroup or workplace that impact the manifestation of skills and behaviors learned in training. According to Noe (1986) the perceived environmental favorability of the workplace will influence the motivation to learn and also to transfer learned skills to the workplace. Noe (1986) conceptualized that there were two components of environmental favorability: a task component related to the availability of equipment and supplies to support trained skills; and a social component which is related to the extent that trainees see opportunities to perform the learned skills in the workplace and the degree of support for the performance of the skills from supervisors and peers.

In terms of the task component of environmental favorability, the perceived availability of resources to support learned skills plays a key role in the transfer of training (Noe, 1986). Peters and O'Connor (1980) identified several categories of perceived work constraints that they believed restricted the use of skills and abilities. These categories include: lack of skills to perform the tasks in the job; lack of services required from co-workers; lack of job-related information; inadequate monetary support; lack of required tools and/or equipment; poor

working conditions; and constrained timeframes in which to perform the work. Further, O'Connor, Peters, Pooyan, Weekley, Frank, and Erenkranz (1984) suggested that perceptions of trainees for such task constraints indirectly influence learning new skills and subsequent behavior change by reducing the motivation to learn or to transfer skills to the workplace.

In terms of the social component of environmental favorability, there are several sources of social support that have been identified in the literature. These are top management, supervisors, peers, and subordinates (see Baldwin & Ford, 1988; Goldstein & Musicante, 1986; Noe, 1986; Noe & Schmitt, 1986). Trainees' perceptions of the level of support from these sources can be seen as playing a key role in the transfer of learned skills. For example, Fleishman (1955), and Hand, Richards, and Slocum (1973) demonstrated that workgroup attitudes played a role in the use of trained skills in human relations on the job. Also, Facteau, Dobbins, Russell, Ladd, and Kudisch (1992) developed Likert-type scales to examine each source of support. Their results indicated that the level of perceived support from subordinates and peers was necessary for training transfer to the work setting. Furthermore, Facteau et al., (1992) demonstrated that supervisors needed to provide opportunities to perform trained tasks by reinforcing the trained skills on the job and by rewarding subordinate and peer support for the use of the skills by trainees. Similarly, Williams, Thayer, & Pond (1991) studied supervisors and managers who had recently completed a rater training program. Their results demonstrated that trainee perceptions of environmental favorability impacted motivation to transfer trained skills to the job, but their model results did not extend motivation to transfer to actual training transfer (Williams et al., 1991).

With this in mind, studies of training effectiveness should include an assessment of the favorability of the post-training environment to support the transfer of trained skills (see Rouiller & Goldstein, 1991; and Tracey, Tannenbaum, & Kavanaugh, 1995). As mentioned in the discussion of implementation quality issues, an assessment of the environment within which learned skills are likely to be performed (organizational analysis) is critical. Without an analysis of the organization and the post-training environment, the capability to assess changes due to training in terms of behavior criteria and in terms of results criteria, will not be possible. What is required is an assessment of trainees' perceptions, beliefs, and/or expectations regarding the post-training environment (Tannenbaum, Mathieu, Salas, & Cannon-Bowers, 1991). Trainee beliefs about the task support of trained skills (e.g., availability of equipment and tools; time to perform tasks; funding to support trained skills; and physical work conditions) should be assessed and reported.

Similarly, the social context of the post-training environment must also be assessed. Trainees' perceptions of the social support are based on the following: opportunities to practice trained skills; reinforcement of performed skills; feedback from peers and supervisors related to skill performance; workgroup cooperation; and organizational climate. The social environment of the workplace will have an impact on the overall effectiveness of training. This is particularly true when considering the use of behavioral and result criteria to evaluate the training program.

In summary, evaluating training programs using behavioral criteria is especially problematic due to the presence of intervening environmental and social variables. These variables reduce the likelihood that positive outcomes from training will be observed in the workplace. Moreover, assessing and accounting for the favorability of the post-training environment as part of a training evaluation study should be useful in explaining the observation of training-related behaviors in the workplace. When accounting for the influence of the post-training environment, this information should be used in the design of the training program to identify potential difficulties in the transfer of training to the job and to develop strategies and approaches for dealing with those difficulties (Thayer & Teachout, 1993; Wexley & Baldwin, 1986). Thus, it is expected that training studies which use behavioral criteria as measures of effectiveness will report lower effect sizes than those that use either learning or reaction criteria. This is possibly due to the influence of environmental and social variables present in the work place. However, studies that account for the post-training environmental favorability and implement strategies to enhance training transfer, in conjunction with the use of behavioral criteria, would be found to be more successful than studies that do not account for the post-training environment. In addition, studies that are found to be effective, in terms of behavioral criteria, would be more likely to report positive outcomes in results criteria terms as well.

Results Measures. Results measures provide an indication of program utility. These measures represent the most distal criteria used to evaluate training effectiveness. Program utility is assessed in terms of the contribution of training to organizational objectives such as lower error rates, lower costs, reduced absenteeism, increased productive capacity, company profits, or workgroup morale (Cascio, 1982). Training evaluations which achieve this level of rigor allow the organization to explore the advantages of training in cost/benefit terms through the use of utility analysis (Cascio, 1989). Utility analysis provides a methodology to assess the dollar value added by engaging in specified training activities.

Given the previous discussion of the role that the post-training environment plays in the manifestation of behavioral changes in the workplace and given that behavioral change is likely to impact on organizational outcome measures (e.g., results), it was expected that studies that

report training effectiveness in terms of behavioral measures were likely to manifest changes in results criteria as well. Thus, it was expected that training programs which were found to be effective in terms of behavioral criteria would also be found to be effective in terms of results criteria, although the number of studies using results criteria was expected to be relatively small.

Finally, the extent to which the task and social dimensions of the post-training environment are maximized might directly influence the manifestation of trained skills in the workplace. Moreover, the actual effectiveness of a training program, in terms of the use of trained skills in the workplace, would be influenced by the trainee's perceptions of the task and social favorability of the post-training environment. That is, as the perceived task and social support for trained skills increases, the magnitude of the effect size for training should increase.

The preceding discussion of Kirkpatrick's topology serves to underscore the problematic nature of using a single criterion measure to evaluate the effectiveness of training. As previously mentioned, Alliger and Janak (1989) noted that there was virtually no relationship between trainee reactions and the other levels. They further noted that there appeared to be a slightly higher correlation among the other levels, but given the small number of studies they identified, the relationship among the levels remains tentative at best.

Future research designed to systematically evaluate the effectiveness of any training or organizational development program must ensure that multiple criteria are used in the evaluation (Tannenbaum & Yukl, 1992). It is important to consider the role that multiple criteria might play in determining the overall effectiveness of training. Given specific training methods, a given group of people, and a known content domain, it is reasonable to ask whether different criteria will provide different information. It was expected that differential outcomes would be obtained from training as a function of the criteria chosen to measure effectiveness. Therefore, the effect associated with the use of different criteria for evaluating training, needed to be quantified.

Training Methods

The choice of particular training methods is another factor that could potentially influence the effectiveness of training. That is, for a given task or training content domain, a given training method may be more effective than others. In a study exploring this issue, Carroll, Paine, and Ivancevich (1972) asked corporate training directors to rate the relative effectiveness of different training techniques for achieving certain training objectives. Their findings indicated that certain training methods were rated as being more effective for specific training objectives (e.g., knowledge acquisition, changing attitudes, problem solving, and interpersonal skill) than others. For example, Carroll et al., (1972) found that lectures were rated as being among the least

effective methods for training across all training objectives. On the other hand, business games were rated as most effective for training problem solving skills.

There are numerous techniques, approaches, and methods which have been used for training in organizations. For convenience and ease of explanation, these techniques can be summarized into two broad categories: (a) on-site methods, and (b) off-site methods (Wexley & Latham, 1991). Table 1 illustrates the methods to be discussed in these categories. While different methods will be discussed within each category, the categories are not to be viewed as mutually exclusive. Each of these categories are next discussed in detail.

TABLE 1

A Classification Scheme for Training Methods

On-Site Training Methods	Off-Site Training Methods
Career development	Lecture
Orientation training	Audiovisual
Job Aids	Programmed instruction
On-the-job training	Computer-assisted instruction
Apprenticeship training	Equipment simulators
Coaching	Teleconferencing
Job Rotation	Corporate classrooms

Adapted from K. N. Wexley & G. P. Latham (1991), *Developing and training human resources in organizations* (2nd ed.). New York: Harper Collins.

On-Site Methods. On-site training methods are those used for providing information and skills to trainees at the work site. The training is conducted within the same physical environment as the actual work to be performed. This tends to enhance transfer of training and reduce the costs associated with training. In addition, many of the techniques to be described in this category actually occur in conjunction with job performance thereby reducing training costs even further.

Numerous types of training methods fall within this category. Career development training is designed to increase the self-awareness and motivation of employees. Orientation training is used to identify and reinforce organizationally "appropriate" behaviors and norms

necessary for advancement within the organization (Cascio, 1991). Job aids are materials which are routinely available to the employee in the work setting which help in the conduct of work. On-the-job training (OJT) is inclusive of a variety of training methods which are conducted during and as part of the work activities. OJT methods can include apprenticeships (mentoring with senior job incumbents), job rotation (working on a variety of related jobs to enhance general expertise across the jobs), and on-the-job coaching (working on job-related tasks with limited monitoring from incumbents). These methods are used for individuals or small groups at different levels within the organization and provide essential basic skills information for the performance of tasks within jobs.

At a general level, Wexley and Latham (1991) have proposed that the on-site methods are used to improve trainee self-awareness, job skills, and motivation. As such, these methods are useful for improving cognitive skills and for enhancing both psychomotor and interpersonal skills in workers. For the present discussion, cognitive skills refer to the thinking, idea generation, understanding, or knowledge requirements of the job. Psychomotor skills refer to behavioral aspects of job tasks and are focused upon the "doing" aspects of work. Interpersonal skills are those related to the interaction of an individual with the workgroup, with supervisors, and with clients or customers. The classification of these specific skills will be more fully explicated in a later section.

With respect to on-site methods, career development training is useful for enhancing cognitive skills such as self awareness, while orientation training and job aids are more useful methods for training job skills. Also, on-the-job training and apprenticeship training are useful for enhancing psychomotor-based job skills. Finally, for interpersonal skills, job rotation and coaching are seen as being useful for enhancing trainee motivation (Wexley & Latham, 1991).

Off-Site Methods. Off-site methods offer a somewhat different approach to training. These methods are employed to provide information to trainees in an environment which is removed from the pressures of the job site (Wexley & Latham, 1991). Lectures, audiovisual techniques, teleconferencing and corporate classrooms attempt to train employees by focusing primarily on the cognitive aspects of job skills. Although lectures and audiovisual techniques can be used in on-site training situations, their primary use is within classrooms away from the workplace. Programmed instruction and computer-assisted instruction are methods which are usually self-paced, sequenced training. In addition to the delivery of training material, these methods usually are capable of tracking the performance of trainees through the instructional sequence. Computer-assisted instruction is usually not used as a stand alone training method, but in conjunction with more traditional approaches such as lectures or classroom training.

Simulation methods can include actual equipment simulations as well as case studies, critical incidents, role playing, in-basket scenarios, and behavior modeling. Simulation methods allow representative organizational problems to be presented to individuals or groups for solution. Simulation techniques are commonly used for leadership and management training. Additionally, these techniques can be used to form a fundamental link between behaviors and solutions for current problems, as well as provide an opportunity to explore strategic problem solving issues.

With respect to off-site training methods, Wexley and Latham, (1991) have proposed that these methods are designed specifically to improve trainee job skills. As such, the methods are useful for improving cognitive skills or enhancing psychomotor skills in workers. For cognitive skills, Wexley and Latham (1991) have proposed that lectures, audiovisual techniques, programmed instruction, teleconferencing, and corporate classrooms are the appropriate training methods. For psychomotor skills, computer-assisted instruction and equipment simulators are the most appropriate methods to use.

An example of the need to select training methods that are appropriate for the type of task trained and the training context can be seen in a recent meta-analysis of flight simulator training effectiveness (Hays, Jacobs, Prince, & Salas, 1992). Hays et al., (1992) found that using simulators, in conjunction with actual equipment (actual aircraft flying), improved training effectiveness and efficiency beyond that obtained by using the actual equipment exclusively. For example, using a flight simulator in conjunction with actual flying time produced improvements above those obtained with actual flying time alone. However, the researchers also demonstrated that the type of task trained and the amount of training provided, heavily influenced training outcomes. When jet-related tasks such as takeoff, approach, and non-carrier landing were used, the training effects were greater than they were for combinations of all tasks (Hays et al., 1992). These results were strongest ($r = .26$) for jet-related tasks. For helicopter tasks, the results were very small ($r = .02$). This may have been due to the fact that there were a very small number of available helicopter training studies ($N = 7$). Furthermore, when trainees were allowed to progress at their own pace (as opposed to group pacing) training effects were improved. In this example, the use of complimentary training methods increased the overall effectiveness of the training program. Using only one or the other method would have resulted in lower overall effectiveness. The choice of training method had a significant impact on the overall effectiveness of training.

In summary, the use of different training methods may influence the effectiveness of training. Certain training methods are more ideally suited to providing training related to a

specific content domain than are others. It is likely that matching the content of the training with an appropriate method to achieve effective training will be extremly important. Thus, it was expected that a given training method would be found to be more effective in certain training situations than in others.

There are several other factors that may influence the effectiveness of training. These include the skills and tasks to be trained, the types of individuals who will receive the training, and evaluation design and methodological issues. The next section explores the potential impact of the type of skill and tasks to be trained on the overall effectiveness of training.

Skill/Task Characteristics

Skill and task characteristics are another factor which may influence the effectiveness of training. That is, the effectiveness of a given training intervention may be a function of the skills and tasks to be trained. For the purpose of training, what must be described are the classes of skill or characteristics of tasks that are trainable. These classes of skill and tasks then, would be the focus of the training activity. In addition to providing descriptions of the classes of skills or tasks to be trained, a general classification scheme of skills and tasks might serve as a basis for specifying training objectives for training related to job functioning. A general classification scheme useful for both skills and tasks includes psychomotor, cognitive, and interpersonal categories (Farina & Wheaton, 1973; Fleishman & Quaintance, 1984; Goldstein, 1993). Psychomotor skills and tasks include the behavioral activities associated with a job. These skills and tasks are related to the hands-on or "doing" parts of the job. Cognitive skills and tasks are related to the thinking, idea generation, understanding, or knowledge requirements of the job. Finally, interpersonal skills and tasks are those which are related to interacting with others in a workgroup or with clients and customers.

It can be argued that in training for complex psychomotor tasks such as repairing subcomponents on a jet engine, using work sample-based techniques such as hands-on, or on-the-job techniques would appear to be more effective than a lecture describing how to remove and replace the subcomponents of the engine. Thus, learning by doing using an actual engine might be more effective. The lecture, then, might not be seen as an effective training method for motor skills. On the other hand, if the goal of the training is to provide knowledge of job content, such as that related to jet engine functioning, aerodynamics, or fluid dynamics, then lectures might be the ideal training method.

Wexley and Latham (1991) have highlighted the need to consider skill and task characteristics in determining the most effective training method. Using the on-site and off-site training method classification and the skill and task categories previously discussed, it is possible

to highlight the relationship between different training methods and appropriate skill and task categories for use (Wexley & Latham, 1991). Table 2 illustrates a proposed linkage among different training methods and skill and task categories.

TABLE 2

A Classification Scheme of Training Methods and Skill/Task Categories

Skill/Task Category	On-Site Training Methods	Off-Site Training Methods
Psychomotor	On-the-job training Apprenticeship training Coaching	Computer-assisted instruction Equipment simulators
Cognitive	Career development Orientation training Job aids	Lecture Audiovisual Programmed instruction Teleconferencing Corporate classrooms
Interpersonal	Job rotation	

Adapted from K. N. Wexley & G. P. Latham (1991), *Developing and training human resources in organizations* (2nd ed.). New York: HarperCollins.

In their study of training director ratings of different training methods, Carroll et al., (1972) noted that the raters rated certain skills (denoted as training objectives) as more effectively trained using specific training methods. For example, knowledge acquisition was rated as being best trained using programmed instruction, while problem-solving skills were rated as best trained using a case study. Similarly, interpersonal skills were rated as being most effectively trained using some form of group exercise or sensitivity training.

Finally, Hays et al,. (1992) found that certain tasks were more effectively trained using a simulator than others. For example, they found that when simulators were used for jet takeoff, landing approach, and landing (non-carrier) the effects of the training were greater than for the combination of all the tasks trained (Hays et al., 1992).

In summary, it is important to adequately match training content to certain training methods, and not to others. Moreover, it is probable that the nature of skills and tasks to be trained will influence the effectiveness of training. Thus, it was expected that a given training method was likely to be more effective for training certain kinds of skills and tasks than for others.

In much the same way that the choice of evaluation criteria, training methods, and the skills and tasks to be trained could potentially influence the effectiveness of training, characteristics that the trainee brings to the training environment may influence the effectiveness of training as well. The impact of these trainee characteristics on the effectiveness of training is explored in the next section.

Trainee Characteristics

Characteristics that the trainee brings to the training situation may potentially influence the effectiveness of training. Research exploring the impact of trainee characteristics has typically focused on selecting individuals who are more likely to be successful in a training program, and less upon matching individuals to training methods that capitalize on certain attributes (Tannenbaum & Yukl, 1992). Moreover, understanding how individual characteristics influence training effectiveness has not been a focus of past research (Tannenbaum & Yukl, 1992). Individuals with markedly different aptitudes, abilities, skills, motivation, self-efficacy, interests, attitudes, and past history are likely to benefit from an instructional intervention in different ways (Katzell & Goldstein, 1989). That is, certain types of instructional interventions will facilitate learning for some individuals and not others.

Aptitude-Treatment Interactions (ATI). An aptitude-treatment interaction occurs when one training method is not viewed as being equally effective for all trainees. An aptitude refers to any measurable characteristic of the individual that is propaedeutic to achievement in a given situation (Corno & Snow, 1986). Aptitude-treatment interaction research (Cronbach & Snow, 1977) attempts to relate measurable characteristics of individuals to the differential effectiveness of various training methods. According to Cronbach and Snow (1977), this relationship can take the form of capitalization, remediation, or compensation. That is, the training method can capitalize upon assets, preferences, or tendencies of the trainee; remediate shortcomings in the trainee; or compensate for trainee weaknesses. Examples of relevant individual difference variables include knowledge, skills, abilities, motivation, age, gender, and attitudes which are brought to the training situation by the trainee.

Considering an ATI approach to training evaluation is a recognition that a single training method may not be equally effective for all trainees. Individuals at different aptitude levels may

perform better (or worse) under different training methods (Corno & Snow, 1985). As an illustration of this, in a study of entry-level auditors, trainees with higher aptitudes in numeric reasoning and general accounting and bookkeeping principles, as measured by content-valid tests of these areas, learned more effectively from on-the-job coaching conducted by a senior auditor. Conversely, trainees with less aptitude in these areas learned most when the coaching was supplemented with lectures and programmed levels of instruction (Wexley & Latham, 1991). This is not to say that individuals lacking the specific aptitudes could not benefit from the training, however, the training program must be tailored to match or account for their level of aptitude. Trainees who had these aptitudes can be seen as being more "trainable" than their lower aptitude counterparts. Additionally, the magnitude of outcomes in terms of the effectiveness of the training, may be larger for higher aptitude trainees than for their lower aptitude counterparts. Therefore, it is conceivable that certain individual aptitude levels in trainees would, in fact, influence the effectiveness of training.

According to Goldstein and Bruxton (1982), trainers recognize the importance of individual characteristics in addition to situational characteristics. However, most instructional designers focus on developing common learning environments that seek to maximize performance of a target set of tasks that must be performed by all trainees (Katzell & Goldstein, 1989).

Empirical research regarding the impact of ATI upon training effectiveness has produced conflicting results. Bracht (1970) reviewed 90 studies of training and found that ATI were found about as often as would be expected by chance. Additionally, Glass (1970) concluded that "there is no evidence for an interaction of curriculum treatments and personological variables. I don't know of another statement that has been confirmed so many times by so many people" (p.210). In a re-analysis of the studies cited by Bracht (1970), and Cronbach and Snow (1977) concluded that the interaction of aptitude and treatment does exist. Christal (1974) and Gettinger and White (1979) found that the rate of skill acquisition was strongly influenced by the learner's preexisting ability level. Wightman and Sistrunk (1987) examined part-task training strategies in a carrier landing simulation training program. Results indicated that trainees with lower motor-skill ability benefitted more (increased training transfer) from a task-segmented (chained) training strategy than from the task simplification strategy. Thus, an aptitude-treatment interaction of motor-ability and training strategy was found. Further, Shute (1992) identified a relationship between associative learning ability and learning as measured by a posttest of declarative electricity knowledge. Using an intelligent tutoring system (ITS) for electricity knowledge, Shute (1992) demonstrated that students who had high associative learning ability were able to

learn more when the ITS required the student to induce principles of electricity. Also, subjects with low associative learning ability learned more when the ITS provided the electricity principle, then required the subject to apply the principle.

Regian and Shute (1991) showed that certain ability groups actually learned better with fewer practice trials. Individuals with high working memory and low general knowledge capacity, learned better in an extended ITS environment that provided skill-specific practice. Conversely, individuals with low working memory and high general knowledge capacity learned better in a more constrained ITS environment that provided minimal (one-third as many) skill-specific practice trials.

Finally, Arthur, Young, Jordan, and Shebilske (in press) explored the potential influence of trainee interaction anxiety on the effectiveness of different training protocols. Their results indicated that for trainees who had high interaction anxiety, an individual training protocol was the most effective training approach. Conversely, trainees with low interaction anxiety performed better as a result of a didactic training protocol.

Research exploring general ability-related ATI have been found to be more consistent than studies involving specific abilities (see Ghiselli, 1973; Lohman & Snow, 1984; Snow & Yallow, 1982; Tyler, 1962). Also, Mumford, Weeks, Harding, and Fleishman (1988) explored the relationship among learner characteristics and course content variables upon training outcomes. Findings indicated that such learner characteristics as aptitude, motivation, and reading grade level were important determinants of training performance (measured by training outcomes) above and beyond the influence of training course content (Mumford et al., 1988).

Aptitude-treatment interactions have been explored across a number of contexts (Ackerman, Sternberg, & Glaser, 1989; Snow, 1986). For example, ATIs have been explored in skill learning and practice (Ackerman, 1987); motivation and cognition (Kanfer & Ackerman, 1989); computer literacy (Wesley, Krockover, & Hicks, 1985); classroom climate (Barclay & DeMeers, 1982); trainee attributions of success and failure (Campbell, 1988); need for achievement and goal orientation (Dweck, 1986; Elliot & Dweck, 1988); and self-efficacy (Bandura, 1986; Bouffard-Bouchard, 1990; Gist, 1989; Gist, Stevens, & Bavetta, 1991; Taylor, Locke, Lee & Gist, 1984).

Motivation has been shown to be an important trainee characteristics for learning and transfer of training. Noe and Schmitt (1986) and Wexley and Latham (1991) have distinguished between ability ("can do") and volition ("will do") to acquire new skills as being centrally important to the effectiveness of training. In addition, Mathieu et al., (1992) explored the use of expectancy motivation measures in a study of clerical workers. Their results indicated that

individuals with higher pre-training motivation were more likely to show improvements as a result of training and to provide more positive reactions to the training when educational and ability differences were controlled (Mathieu et al., 1992).

Although the impact of ATIs have been extensively studied in a variety of settings, they have not been studied to a great extent within organizational settings (Tannenbaum & Yukl, 1992). One potential area for research in the organizational arena is related to general academic ability and the complexity of the instructional program (Tannenbaum & Yukl, 1992). Lohman and Snow (1984) demonstrated that high ability students benefitted more from training programs which were less structured (e.g., emphasizing independent knowledge acquisition) and more complex. With this in mind, it is important to assess trainee's prior knowledge and achievement and tailor the training program to capitalize on this experience and achievement (Campbell, 1988). One benefit from this assessment is to identify how the training might be sequenced in a manner that is consistent with the requirements of the job and with the capabilities of the trainee (Katzell & Goldstein, 1989). Second, determining the rate of acquisition and the duration or survivability of certain training content and relating these rates to assessed trainee capabilities will allow the specification of appropriate retraining intervals to maintain skill performance (Ackerman, 1987; Kanfer & Ackerman, 1989).

Thus, the empirical evidence from ATIs research, although somewhat mixed, generally supports the existence of aptitude-treatment interaction effects. Several explanations for the inconsistencies seen in the research have been proposed. ATIs studies conducted prior to 1977 typically had 40 or fewer subjects per condition and, therefore, lacked sufficient statistical power to detect significant effects (Cohen, 1988; Schmidt & Hunter, 1978). The studies did not control for plausible alternative explanations of the observed effects. For example, training time was allowed to vary along with achievement. Finally, ATI studies generally did not randomly assign subjects to conditions or use experimental or quasi-experimental designs.

In summarizing their findings on ATI, Cronbach and Snow (1977) provide several recommendations for future ATI research designs. These include (a) using larger sample sizes; (b) randomly assigning subjects to conditions; and (c) holding training time constant and allowing achievement to vary. Given the potential moderating influence of trainee characteristics on training outcomes, ATI should be considered and evaluated in any assessment of training effectiveness.

The traditional approach to the design and conduct of education and training activities has taken the position that one training program is equally effective for all trainees. In terms of efficiently providing a baseline level of knowledge and information to a large and diverse

population of trainees, this may be a reasonable approach. However, in terms of maximizing individual learner's capabilities to derive salient and useful information from the learning situation, a more individualized approach may be required. Given the somewhat mixed findings related to ATI and the recent research evidence outlining the impact of other individual characteristics on the effectiveness of training, our present understanding and research evidence regarding the impact of different training approaches for varying content domains and different individuals is, at best, tentative. Campbell (1988) has suggested that trainee achievement and experience may interact with training complexity or difficulty. He further noted that, if this is true, training developers must pay more attention to the measurement of trainee achievement and experience and match the training program to the capabilities of these trainees. With this in mind, the impact of ATI on organizational training effectiveness needs to be explored in future research.

Given the debate over the existence of ATI, the importance or non-importance of considering trainee characteristics and their role evaluating the effectiveness of training, and the divergence of findings in the empirical literature with respect to ATI, this study used a meta-analytic approach in an attempt to explore the issue and potentially contribute to the resolution of this debate. That is, would studies that account for trainee differences in the training program report larger effect sizes than those that do not consider trainee differences? Or would the converse be found?

Study Design and Methodological Issues

A final factor that may influence training effectiveness is the rigor of the evaluation study. The level of control exercised in empirical studies of training effectiveness can influence the conclusions related to the efficacy of different training methods. Designing and implementing rigorous studies of training effectiveness is extremely difficult in applied settings. The paucity of rigorous empirical studies evaluating training is evidence of these difficulties (Campbell, 1971; Goldstein, 1980; Wexley, 1984). Typically, the "best" approaches should use experimental or quasi-experimental methods to systematically explore the linkage of change to training (Arvey & Cole, 1989). Rigorous control enables researchers to identify and distinguish true change at all levels (alpha change) from changes in scale recalibration (beta change) and changes in the trainee's conceptualization of the construct (gamma change) (Terborg, Howard, & Maxwell, 1980). To ensure that the findings observed across a body of literature are attributable to the intervention and not to such influences as experimenter expectancies, researchers must be sensitive to issues associated with the methodological rigor in the design of evaluation studies (Barrick & Alexander, 1987; Terpstra, 1981; Woodman & Wayne; 1985).

Terpstra (1981) demonstrated the existence of a positive-findings bias which was attributable to the methodological rigor of organizational development (OD) intervention evaluation studies. Positive-findings bias is the tendency of reported results in evaluation studies to be more positive if the methodological rigor of the design used in the study is poor. Thus, as methodological rigor increases, the likelihood of obtaining positive findings decreases. Terpstra (1981) found OD intervention studies reporting uniformly negative results were more rigorous in terms of their methodology and design and conversely, studies which reported uniformly higher positive results were found to be the least rigorous. Other studies in Management by Objectives (MBO) effectiveness (Kondrasuk, 1981), and psychologically based interventions (Guzzo, Jette, & Katzell, 1985) have obtained similar positive-findings bias results.

In a subsequent study of the efficacy of a specific OD intervention, quality circles, Barrick and Alexander (1987) obtained results that were markedly different from those observed by Terpstra. Barrick and Alexander (1987) found that differences in observed methodology/design scores were not significant. This was the case across studies reporting negative, mixed, and positive outcomes related to quality circles. Their results were similar to results found in two other studies of OD evaluation (see Bullock & Svyantek, 1983; Woodman & Wayne, 1985). Barrick and Alexander (1987) proposed that one possible explanation for positive-findings bias in some studies and not in others, might be due to the inclusion of "popular press" studies. Typically, the quality or design rigor of these studies limits their empirical utility. This is due to the fact that there is a significant lack of experimental control in the design of these studies and this lack of control makes it virtually impossible to distinguish between experimenter expectancies and true experimental differences between groups in the study.

In a recent study exploring positive-findings bias in OD interventions, Roberts and Robertson (1992) found that the use of different combinations of methodological criteria produced different results, but were generally consistent with those of other recent studies (e.g., Barrick & Alexander, 1987; Bullock & Svyantek, 1983; Woodman & Wayne, 1985). In addition, Roberts and Robertson (1992) found only one instance where positive-findings bias was evidenced - when sampling criteria were used to evaluate methodological rigor. While the positive-findings bias issue remains unresolved, the authors cautioned that the composition of criteria used to evaluate methodological rigor can significantly impact study outcomes (Roberts & Robertson, 1992).

Because training is similar to OD types of interventions, coupled with the equivocability of past research findings, the present study explored the existence of a positive-findings bias effect in the training evaluation literature. Terpstra (1981) identified several methodological

characteristics of evaluation studies that need to be considered. To quantify these study characteristics, Terpstra (1981) developed a methodological rigor scale to be used as a means of "grading" the design or methodological rigor of empirical studies. Woodman and Wayne (1985) suggest a modified method for scoring studies. This method evaluates studies in terms of methodological and design characteristics based upon Terpstra's (1981) approach and a nine dimension scale developed by Woodman and Wayne (1985). The modified method proposed by Woodman and Wayne (1985) evaluates design considerations related to (1) representative sampling strategy, (2) sample size, (3) utilization of control groups, (4) random assignment, (5) repeated measures strategy, (6) reliability and validity information, (7) significance level, (8) inclusion of objective data (organizational level information), and (9) the use of multivariate analysis procedures. Methodological shortcomings and problems with reporting critical information in evaluation studies have been a consistent problem in the training literature (Arvey, Cole, Hazucha, & Hartano, 1985; Burke & Day, 1986). Therefore, methodological considerations are seen as essential for assessing the effectiveness of training. However, given that recent results related to positive-findings bias are somewhat mixed, the impact of the design and/or methodological rigor of the evaluation study on training outcomes remains questionable.

In summary, the methodological rigor of a training evaluation study may impact the overall effectiveness of the training program. Thus, based on the mixed findings from past research, studies that are methodologically more rigorous might, or might not, be less likely to obtain positive training outcomes. Conversely, those studies which are less rigorous may, or may not, obtain more positive training outcomes due to the lack of experimental control. On the other hand, by accounting for the methodological rigor of the study design, any differences in the reported effect sizes should be obtainable with meta-analysis techniques, and therefore provide evidence to help resolve the debate over the impact of methodological rigor on training effectiveness.

Meta-Analysis Techniques

The present study used meta-analytic procedures (see Glass, McGaw, & Smith, 1981) to assess the influence of the identified factors on training effectiveness. Meta-analysis is a statistical tool for summarizing empirical results across a number of studies to reach a quantitative generalization. Although there are a number of meta-analytic approaches and techniques (Bangert-Drowns, 1986; Glass et al., 1981; Hunter & Schmidt, 1990; Hunter, Schmidt, & Jackson, 1982), the basic goal of each approach is conceptually the same - cumulating results from several primary studies within a content domain into an overall quantitative summary.

Meta-analysis techniques provide an alternative to the more traditional methods of summarizing results across a body of literature. The traditional method involves a narrative review, where the reviewer arrives at nonquantitative conclusions based upon an exhaustive reading of the literature. This more traditional method has been criticized as being somewhat weak as a means of integrating a body of literature (e.g., Glass et al., 1981). Finally, qualitative approaches are not sensitive to subtle statistical results and findings (Green & Hall, 1984).

In contrast, there are several methods for summarizing research results in a more quantitative manner. A minimal quantification technique involves significance "vote-counting" or box scoring, where physical counts of the number of significant results favoring a hypothesis, the number of significant disconfirming results, and the number of nonsignificant results are made. A variant on the "vote-counting" approach involves tallying the study results for particular groups (e.g., men and women) and examining the direction of the effect. Also, study results can be examined using both the level of significance and the actual probability of a chance occurrence of the null hypothesis (*p*-value) (Green & Hall, 1984, p. 42).

As a more quantitative and standardized approach, meta-analysis techniques offer several advantages over other approaches to aggregation. First, meta-analysis is an efficient approach for quantifying the results of a large volume of literature. Second, meta-analysis, like other quantitative approaches, is seen as relatively more objective than narrative reviews. Third, meta-analytic approaches can detect relationships and trends that may be too subtle to be detected in the narrative review and other approaches. Fourth, meta-analysis approaches allow researchers to explore main effects and interactions across studies. Fifth, it is possible for the meta-analytic researcher to test hypotheses that were never tested in the original studies. Sixth, meta-analysis techniques can be used to highlight shortcomings and gaps in the existing literature (see Green & Hall, 1984; Guzzo, Jackson, & Katzell, 1989). Finally, Schmidt (1992) has pointed out that meta-analysis techniques are not strictly a means for reviewing literature, but are also a "new way of thinking about the meaning of data" (p. 1173). Further, meta-analysis techniques permit hypothesis testing and addressing research issues that would not otherwise be possible without collapsing or aggregating across multiple primary studies. As discussed in earlier sections, the present study used meta-analysis techniques to help explore issues surrounding some longstanding debates in the training community by cumulating findings based on primary research in the extent literature.

In addition to the advantages previously highlighted, there are several cautions to be considered when using any meta-analysis technique. For example, a study must report sample sizes along with other pertinent (and descriptive) information. This information includes

statistics that allow for the computation of an effect size statistic (e.g., group means and standard deviations). For studies that report statistics such as correlations, univariate *F*, *t*, χ^2, etc., these can be converted using the appropriate conversion formulas (see Glass et al., 1981).

Also, meta-analysts must make several judgment calls when implementing a meta-analysis. The effects of these judgment calls have been shown to account for differences in ostensibly objective meta-analytic studies of the same content area (for a review of some of these studies, see Wanous, Sullivan, & Malinak, 1989). Specifically, the following steps in the implementation sequence, have been noted to call for some judgment on the part of the meta-analyst (e.g., Abrami, Cohen, & d'Apollonia, 1988; Wanous et al., 1989): (1) topic selection - defining the research domain; (2) specifying the inclusion criteria; (3) searching for and locating relevant studies; (4) sampling and selecting the final set of studies; (5) extracting data and coding study characteristics; (6) deciding to group or separate multiple measures of independent and dependent variables; and (7) selecting potential moderators.

In addition, Arthur, Bennett, and Huffcutt (1994) have proposed that the *data analysis* step in the implementation sequence should also be considered as one that calls for some judgment and a decision on the part of the researcher. Unlike other widely used statistical techniques such as *t*-tests, analysis of variance, and measures of association (e.g., Pearson's *r*), which are readily available in statistical software packages, the researcher has to make a decision as to how the data analysis (i.e., calculating mean correlations and correcting for artifacts) will proceed when conducting a meta-analysis. These choices range from using one of several available programs without any modifications, modifying these programs, writing one's own program based on the available correction formulas, to maybe even doing some or all the calculations by hand.

Fortunately, in terms of the software and programs compared by Arthur et al., (1994), it would seem that the choice of which software or program to use is one decision in the implementation of a meta-analytic study that does not have a major impact on the study's outcomes. While there were some differences in the values obtained, they tended to be relatively small. However, future meta-analytic studies should report the specific data analysis programs and procedures used an integral part of their methodology (Arthur et al., 1994).

So, when conducting a meta-analysis, it is crucial to document the steps (e.g., the judgment calls made), taken during implementation. Documenting these judgment calls allows other researchers to evaluate the results in terms of the assumptions made, and to determine the extent to which the obtained results are indicative of the research in the content domain.

Usually, the existence of one or more moderator variables in meta-analysis is based upon some theoretical foundation and is typically hypothesized and coded in the analysis (Hunter & Schmidt, 1990; Steiner, Lane, Dobbins, Schnur, & McConnell, 1991). Statistically, the presence of moderators is also indicated when a substantial amount of unexplained variance remains (Hunter & Schmidt, 1990). Analyses of moderator effects are conducted by separating the dataset into subsets according to the various levels of the potential moderator variable. The mean effect size and variance are then recalculated separately for each level of the variable. The existence of the moderator variable is generally confirmed if the means effect size of the subgroups differ and the variance within each group is less than the original variance.

In summary, meta-analysis techniques offer numerous advantages over more traditional narrative reviews. However, it is important to note that qualitative reviews provide important information as well.

Summary

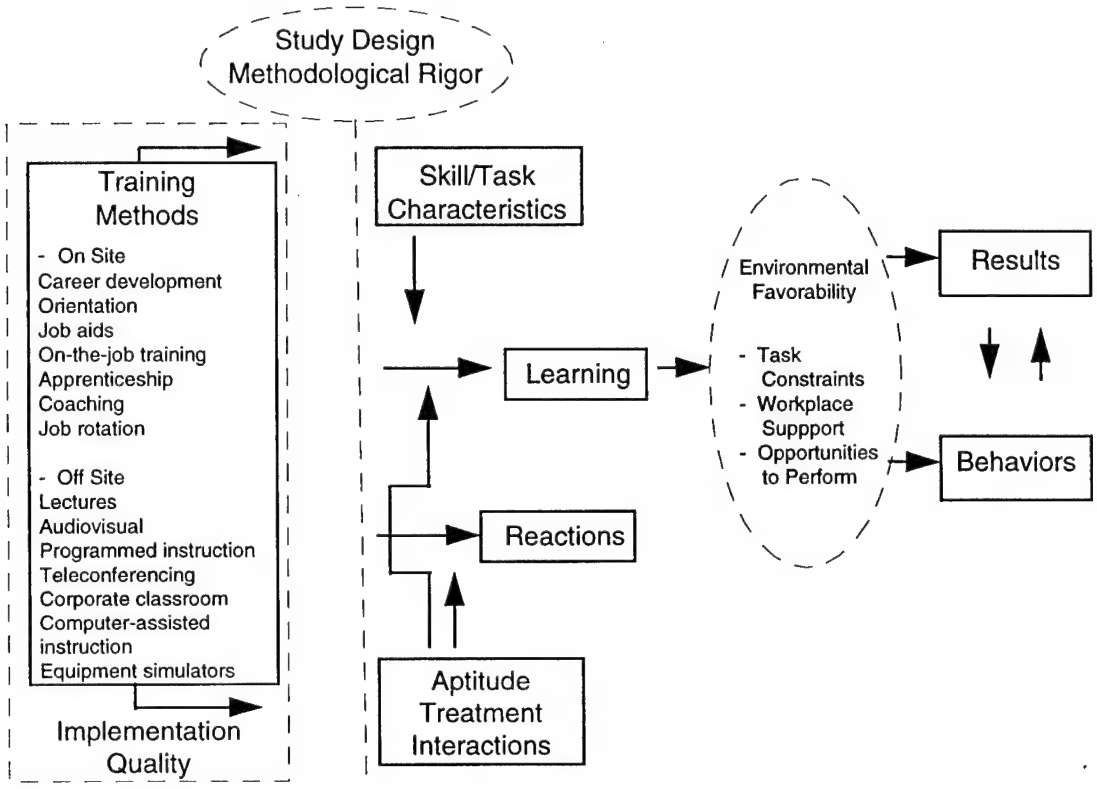
The purpose of the preceding review was to identify several factors that could potentially impact training effectiveness and to discuss *how* they could do this. These factors are: (a) implementation quality; (b) criterion measures of training effectiveness; (c) different training methods; (d) the match between training methods and skill and task characteristics; (e) trainee characteristics; and (f) the methodological rigor/empirical study design of training effectiveness studies. The importance of each of these factors was discussed.

THE PRESENT STUDY

Based upon the issues and empirical research discussed in previous sections of this report, a conceptual framework of training effectiveness was developed. This framework, shown in Figure 1, depicts training effectiveness at a level where the effect of the proposed factors can be evaluated.

FIGURE 1

A Conceptual Framework of Factors That Potentially Influence Training Effectiveness



Components of the Conceptual Framework

Several factors that potentially influence effectiveness have been proposed in this report. The following sections detail the steps that were taken to quantitatively cumulate the training literature using meta-analysis in an attempt to quantify the effect of each factor on training effectiveness. The favorability of the post-training environment for the transfer of training is included as a separate "block" in the framework since it is considered to influence the measurement of behavioral and organizational (results) criteria. Each factor discussed in this

training. The components of this conceptual framework are described in the following paragraphs.

The conceptual framework shown in Figure 1 provides a convenient heuristic within which several hypotheses regarding the nature of the influence of these factors on the effectiveness of training can be formulated and explored. Implementation quality has been shown to be a critical first step in the development, conduct, and evaluation of training. With this in mind, implementation quality is shown as encompassing training methods used to deliver training. In fact, implementation quality could be seen as encompassing the entire framework. However, for ease of explanation and understanding, it is illustrated as a first step in the framework.

Specific training methods are also identified in the framework. Different training methods can be used to deliver different skill and task information to different trainees. The information delivered by different methods is directly effected by the strategies used in the implementation (e.g., person, task, organizational analysis) and is designed to communicate specific skill and task information to a pre-specified group of trainees.

In a manner similar to implementation quality, the design or methodological rigor of the training study plays a key role in filtering the effectiveness of training. That is, the reported effectiveness of training, as measured by any outcome measure, is dependent upon the design of the evaluation study. Training methods are used to deliver different training content to different trainees and the combination of these impacts the effectiveness of training as measured by different types of evaluation criteria (e.g., reactions, learning, behaviors, and results).

Initially, the training content, delivered by specific methods to identified individuals, can be seen as impacting one and possibly two types of criteria, namely reactions and learning. However, note that in the figure, reaction criteria are not linked to learning criteria. As discussed in the review of the literature, the linkage between reaction criteria and learning has not been consistently found (e.g., Alliger & Janak, 1989; Mathieu et al., 1992). Reaction and learning criteria are depicted as being closer, or more proximal, to the delivery of training in a manner similar to how they are likely to occur in the actual training study.

Outcomes of training may or may not be manifest in subsequent job behaviors and organizational indicators as a function of the favorability of the post-training environment for the performance of the learned skills. That is, what is learned may not be transferred to the job in the form of behaviors due to the favorability or unfavorability of the post-training environment. This is evident in the framework where the outcomes from learning criteria are impacted by post-training environmental favorability. Change resulting from training is less likely to be seen in the

more distal types of criteria due to environmental constraints. Finally, results and behaviors are shown to be interrelated. That is, changes in behaviors resultant from a training program may lead to changes in other types of organizational criteria such as absenteeism, turnover, and promotion.

The goal of the present study was therefore to address a series of research questions to be described, using an effect size (ES) meta-analysis approach (Glass et al., 1981; Hunter & Schmidt, 1990; Rosenthal, 1978). The intent was not to merely identify whether a given factor influences training outcomes, but to also provide an indication of the quantitative impact of each factor.

Statement of Hypotheses

A number of specific hypotheses were addressed in this study. First, what is the overall effectiveness of training? That is, aggregating across all factors, what is the global effect size (ES) for training? Is the effect size non-zero?

H1: Training will have a positive overall ES. When corrected, this ES was expected to be fairly small. In addition, the magnitude of the associated standard deviation is expected to be large enough to indicate the presence of one or more moderators.

As suggested in the preceding sections, there are several factors that could potentially influence the effectiveness of training interventions. One such factor is the quality of the implementation process. This factor is seen as important due to the fact that a systematic needs assessment provides the mechanism whereby the questions essential to successful training programs can be identified and addressed. Assessing the training needs of the organization in 0 terms of identifying job requirements to be trained, who needs training and the kind of training to be delivered, should result in more effective training programs. Therefore, studies which report a comprehensive needs assessment (e.g., person, task, and organizational analysis) are seen as being of higher quality in terms of their implementation. It was expected that studies that are of higher quality, in terms of their implementation, as measured by reported needs assessment, should result in more positive outcomes from training. It was therefore hypothesized that:

H2: Training studies with higher implementation quality would result in a larger effect size (that is, more effective training) than those of lower implementation quality.

In terms of criteria used to evaluate training, four types of criteria were discussed: reactions, learning, behaviors, and results (Kirkpatrick, 1959; 1987). It was expected that differential outcomes will be obtained from training as a function of the criteria chosen to measure effectiveness. Reaction measures have been described as the most proximal criteria to use for evaluating training outcomes. As such, reaction measures are minimally impacted by

such things as the organizational environment, resource availability, and group-characteristic bias noted in the earlier discussion of criteria development. However, reaction measures may be deficient as criteria since they typically do not reflect components of performance that are related to the job.

Learning measures are more proximal criteria than either behaviors or results criteria, as measures of training effectiveness. As such, learning measures are also less susceptible to organizational intervening variables such as the social environment of the workplace or supervisory support of trained tasks. Also, learning criteria are usually easy to obtain and fairly nonintrusive in terms of their impact on work activities since they are obtained at the end of the training activities.

The use of behavioral criteria as measures of training effectiveness is especially problematic due to the impact of environmental and social variables. These variables reduce the likelihood that positive outcomes from training will be observed in the workplace. Thus, it was hypothesized that:

H3: Training programs which use reaction measures as criteria would report larger effect sizes than those that use other, more distal criteria such as learning, behavior, and results. In addition, training programs which used learning criteria as measures of effectiveness, would report larger effect sizes than those that use either behavior or results criteria. Further, training studies which used behavioral criteria as measures of effectiveness, would report smaller effect sizes that than those that used either learning or reaction criteria. Lastly, studies that used results criteria would report smaller overall effect sizes than those for reactions, learning, or behavior criteria.

Accounting for the favorability of the post-training environment in the conduct of a training study and as part of a training evaluation study, should be useful in explaining the manifestation of training-related behaviors in the workplace.

Therefore, it was hypothesized that:

H4: Studies that accounted for the post-training environmental favorability in conjunction with behavioral criteria would report higher effect sizes (that is, be found to be more successful) than studies that do not account for the post-training environment favorability.

Also, it was expected that:

H5: Training programs which were found to be effective in terms of behavioral criteria would also be found to be effective in terms of results criteria, although the number of studies using results criteria was expected to be relatively small.

With respect to the use of different training methods, different methods may influence the effectiveness of training. Certain training methods are more readily suited to providing training related to specific content than are others. It is important to match the content of the training with an appropriate method to achieve effective training. Similarly, the characteristics of the skills and tasks to be trained are also factors that could potentially influence the effectiveness of training. It is probable that the nature of skills and tasks to be trained will influence the effectiveness of training.

Thus, it was hypothesized that:

H6: A given training method would report a larger effect size when used to train certain skills and tasks than when used to train others.

In addition, issues related to the existence of ATI and the impact of methodological rigor on training effectiveness were examined in this study. Although no formal directional hypotheses were postulated for trainee characteristics (ATI) and methodological rigor, these variables were examined from an exploratory perspective. That is, it was hoped that evidence from the meta-analysis might contribute to the resolution of the debates and the divergence in the primary studies related to these variables.

METHOD

Literature Search

The present study reviewed the published training and development literature from 1960 to 1993. This time frame was chosen because it encompasses a rather comprehensive training evaluation literature where technological and methodological advances impacted the development, delivery, and evaluation of training. In particular, the period from 1980 to 1993 is marked by improvements in needs assessment methods, increased technological sophistication in training methodologies, and the use of more comprehensive training evaluation techniques and statistical approaches. The increased focus on quantitative methods for the measurement of training effectiveness is critical for a quantitative review such as the one accomplished in this study. However, similar to past training and development reviews (e.g., Latham, 1988; Tannenbaum & Yukl, 1991; Wexley, 1984), the present only included the practitioner-oriented literature if those studies met the criteria for inclusion as outlined below. The present study, therefore, used scientific studies published in journals and books/book chapters which were related to the evaluation of a organizational training program or those which measured some aspect of training effectiveness.

An extensive literature search was conducted to identify empirical studies that involved an evaluation of a training program or measured some aspects of training effectiveness. This search process started with a search of nine (9) computer databases (*Defense Technical Information Center [DTIC], Econlit, Educational Research Information Center [ERIC], Government Printing Office [GPO], National Technical Information Service [NTIS], PsychLit, Social Citations Index [SSCI], Sociofile* and *Wilson*) using the following key words: training effectiveness, training evaluation, training needs assessment, training efficiency, and training transfer. The electronic search was supplemented with a manual search of the reference lists from the past qualitative reviews of the training literature (e.g., Campbell, 1971; Goldstein, 1980; Latham, 1988; Tannenbaum & Yukl, 1992; Wexley, 1984). Approximately 3600 citations were obtained as a result of this initial search. A review of the abstracts of these citations for appropriate content (i.e., empirical studies that actually evaluated a training program or measured some aspect of training effectiveness), along with a decision to retain only English language articles, narrowed the list down to 342 articles. In addition, the reference lists of these articles were reviewed. As a result of these efforts, an additional 253 articles were identified, resulting in a total of 595 articles. This total represents all identified articles. As such, it includes articles representing all types of training programs (e.g., clinical, educational, organizational and rater). Each article was then reviewed and considered for inclusion in the meta-analysis. The sources of

the reviewed articles were as follows: journal articles (97%), books/book chapters (2%), and peer-reviewed conference papers and presentations (1%).

Study (Datapoint) Inclusion Criteria

A number of decision rules were used to determine which studies would be included or retained for the meta-analysis. First, to be included in a meta-analysis, a study must have investigated the effectiveness of a training program or conducted an empirical evaluation of a training method or approach. Studies that reported empirical data on effectiveness or efficiency of a program using single or multiple criteria or measures were also included. Second, studies evaluating the effectiveness of rater training programs were excluded. These studies were excluded because they were considered to be qualitatively different from the more traditional organizational training study or program. Specifically, rater training was not considered to be an intervention that impacts "organizational-related" tasks or activities. Third, to be included, studies had to report sample sizes along with other pertinent (and descriptive) information. This information included statistics that allowed for the computation of a *d* statistic (e.g., group means and standard deviations). For studies that reported statistics such as correlations, univariate *F, t*, χ^2, etc., these were converted to *d*s using the appropriate conversion formulas (see Glass et al., 1981; Hunter & Schmidt, 1990; Wolf, 1986). For studies that contained incomplete data required for the meta-analysis, an attempt was made to contact the primary authors to obtain the additional information.

Data Set

Non-independence. As a result of the inclusion criteria an initial data set of 1219 data points (*d*s) from 177 articles was obtained. However, many of the data points were non-independent. Effect sizes or data points are non-independent if they are computed from data collected on a single group of subjects. Decisions about non-independence have to also take into account whether the effect sizes represent the same variable/construct or not. The effect of non-independence is to reduce the observed variability of the effect sizes. Thus under these conditions, interpretations of the homogeneity of effect sizes (e.g., chi-square tests) must be made very cautiously. Another effect of non-independence is to artificially inflate sample sizes and effects beyond the number of independent data points. Although this may increase the power of the meta-analysis, it becomes difficult to determine the amount of error in the statistics describing the data points. A final effect of non-independence is to over weight the contribution (either positively or negatively) of the studies/articles contributing multiple non-independent data points.

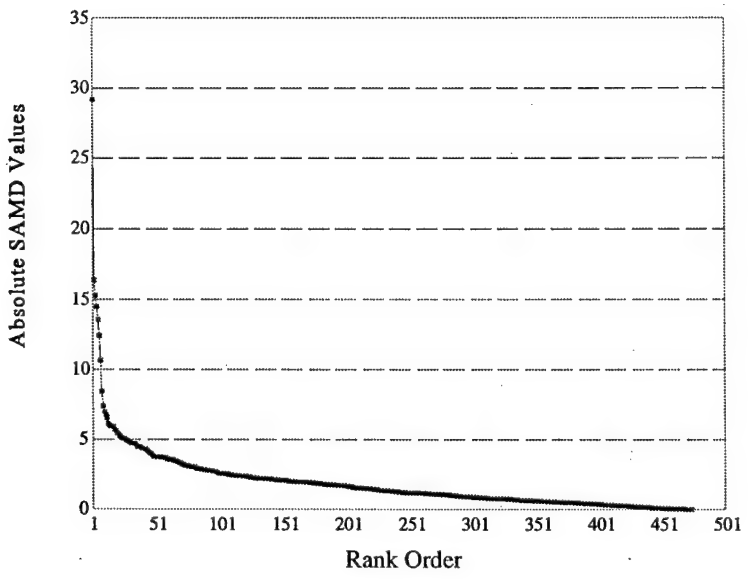
Consequently, to address these problems, when data points are non-independent, the accepted practice is to aggregate them by finding the average. Implementing this practice resulted in 474 independent data points from the 177 articles.

Identifying Outlier Data Points. Huffcutt and Arthur's (1995) sample-adjusted meta-analytic deviancy (SAMD) statistic was next computed for each data point to detect outliers. In Huffcutt and Arthur's (1995) procedure, outliers or extreme data points are identified using a scree plot (Dillon & Goldstein, 1984; Loehlin, 1987) to set a cut-off (the scree) above which data points are considered to be outliers. Specifically, the absolute values of the SAMD statistics are rank ordered from the highest to the lowest and plotted. SAMD values which rise above the flat gradual slopes are identified as potential outliers and investigated.

The SAMD statistics were computed for each data point. Across all 474 *d*s, the mean SAMD value was .0666 (*SD* = 3.032). The resulting SAMD scree plot is presented in Figure 2.

FIGURE 2

Scree Plot of Absolute SAMD Values for Each Data Point



As this chart presents, the first eight data points appear to rise above the flat portion of the plot and thus were identified as outliers. The cut-off value was 8.445 with absolute SAMD values ranging from 8.445 to 29.154 for the eight *d*s. These eight *d*s constituted 1.69% of the

474 *d*s in the data set. Table 3 presents the absolute *d* and absolute SAMD values for each of the eight outlier data points.

TABLE 3

Absolute d and SAMD Values for Each Outlier Data Point

Data Point	Absolute *d*	Absolute SAMD
1	7.514	29.154
2	4.533	16.407
3	4.150	15.305
4	5.253	14.492
5	3.868	13.563
6	3.618	12.451
7	4.259	10.667
8	3.232	8.445

Dropping the eight outliers resulted in a final data set of 466 independent *d*s. The sources of these data points were as follows; journal articles (99.8%) and books/book chapters (.2%). A reference list of articles included in the final data set for this meta-analysis is provided in Appendix A.

General Coding Procedures

Each article was coded for basic information. This information included identifying and recording the descriptive (e.g., mean and standard deviation) and inferential statistics (e.g., *F*, or *t*-tests) the source of the study (e.g., journal article, book/book chapter, conference paper/presentation, or technical report), and the study type (e.g., basic/laboratory study or applied/real world/field study).

Empirical studies were coded with respect to the variables outlined in conceptual framework previously discussed. Detailed descriptions of the coding procedure used for each variable are provided in a subsequent portion of this section. The actual coding data sheet which accompanied each study is included as Appendix B. Each study was coded in terms of the method used to deliver training. In addition, the effectiveness measure used in each study was coded. Studies were also coded into categories related to the quality of implementation, type of task or job content being trained, presence or absence of ATI, study design or methodological rigor of the reported study, and the presence of indicators of post-training environmental favorability.

Description of Variables

This section presents a description of the variables that were coded for the meta-analysis.

Implementation Quality. Implementation quality was coded as the extent to which a systematic needs assessment was conducted and reported in each study as part of the training development activity. Separate codes for each of the three aspects of a needs assessment (e.g., organization, task, and person analysis) were assigned. A code of 1 was assigned if the particular aspect of needs analysis was addressed, 0 otherwise. Therefore, the maximum study implementation quality "score" that could be obtained would be 3, indicating the reported presence and use of all these aspects of needs assessment.

Training Method. Each of the training methods used in each study was coded. In addition, if a training method was used in the study that was not in the list of methods provided, that method was coded in the "Other" space on the coding sheet. A method was coded 1 if it was the method used to deliver training in the study, 0 if that method was not used.

Skill/Task Characteristics. The training content which was the focus of each identified study was coded. For example, if the focus of the training program was to train psychomotor skills and tasks, then the psychomotor characteristic received a 1 while the other characteristics (e.g., cognitive, interpersonal, or other) were coded as 0's. Again, the purpose of the coding was to assign a unique code to each of the characteristics.

Trainee Characteristics. Coding for trainee characteristics was done at two levels. Initially, each study was coded in terms of whether it addressed, or failed to address trainee differences as defined in the ATI literature. Studies that accounted for trainee differences were assigned a 1 while studies that did not consider ATI received a 0 code. For those that did address ATI, a second level of coding was conducted to identify the specific trainee characteristic(s) that were considered and addressed in the study. Again, the coding scheme was such that each characteristic would receive a unique code. The list of coded characteristics included the following: education, aptitude, experience, motivation, self-efficacy, personality, locus-of-control, cognitive ability, psychomotor ability, gender, ethnicity, age, SES, and other. In some cases, more than one characteristic was identified. In these instances, each characteristic was identified and coded.

Study Design/Methodological Rigor. The level of rigor or quality associated with each study was assessed using the methodological rigor scale outlined in Terpstra (1981). Terpstra (1981) developed and has applied this scale to evaluate and rate the design/methodological rigor of research studies. The scale proposed by Terpstra (1981) includes the following ratings which were assigned to each study in the meta analysis: (1) Sampling strategy - if a probabilistic

sampling strategy (e.g., stratified, cluster, random) was reported, the study received a 1, else 0; (2) Representative sample size - $N \geq 30 = 1$, $N \leq 30 = 0$; (3) Presence of control group; (4) Random assignment to conditions - 1 if one or both were present, else 0; (5) Pretest/posttest design - 1 = yes, 0 = no; (6) Cutoff for significance - $\alpha \leq .05$ or less = 1, else = 0.

In addition, the three items identified by Woodman and Wayne (1985) were also included and coded in the present study. These three items were (7) a reported reliability coefficient $\geq .60$ for the dependent variable and/or some indication of the validity of the dependent variable - 1 if reported, 0 if not reported; (8) objective criteria for dependent variables (e.g., results criteria reported) - 1 if reported, 0 if not reported; and (9) use of a multivariate analysis procedure - 1 if reported, 0 if not reported.

Studies were coded for methodological rigor using these 9 criteria to obtain a total methodological grade or score for each study (Barrick & Alexander, 1987; Roberts & Robertson, 1992; Terpstra, 1981; Woodman & Wayne, 1985). This grade or score was used as the indicator of study design/methodological quality in this meta-analysis. The maximum grade obtainable was 9. A grade or score of 9 would indicate the presence of all methodological rigor criteria. However, no studies coded for this meta-analysis obtained the maximum grade or score.

Evaluation Criteria. Evaluation criteria types which were used in each study were coded as 1 if used or 0 if not used.

Environmental Favorability. Environmental favorability has been discussed in this report as potentially impacting the measurement of criteria used to evaluate training effectiveness. With this in mind, studies that accounted for environmental constraints and addressed these within the training program were coded for inclusion in the following manner. Environmental favorability has been discussed in the literature as being composed of two major dimensions, a task dimension and a social dimension. Thus, coding for environmental favorability was done at two levels. First, each study was coded to indicate whether it considered and assessed any post-training environmental variables. Those studies which addressed post-training environmental characteristics in the conduct of the training program received a 1, while those that did not received a 0. For those studies that received a 1, a second level of coding was conducted. Coding was accomplished for both task and social dimensions of favorability where applicable. This included separate coding for reported indices of the task dimension (e.g., availability of tools/equipment, availability of supplies, familiarity with task, availability of monetary resources, time to perform, post-training environment working conditions, or an other category) and for reported measures of the social dimension of favorability (e.g., top management support, supervisory support, peer support, subordinate support, opportunities to perform, feedback,

reinforcement, and an "other" category). Unique codes for each indicator or measure used, and the mean reported rating derived from each measure were recorded.

However, this particular variable was not analyzed in the meta-analysis. This was because the dataset did not contain any studies reporting the consideration of environmental favorability. The assessment of the favorability of the post-training environment is a relatively recent issue. As such, no studies were available to code. It is expected that future empirical studies will report assessments of the post-training environment.

Coding Procedures and Interrater Agreement

The present study used procedures developed and tested in another meta-analysis (see Arthur, Bennett, Stanush, & McNelly, 1995b) to train coders. These procedures were also used to assess the accuracy and degree of agreement between the author and a industrial/organizational psychology graduate student who also coded the studies. The first author of the present study and graduate student will subsequently be referred to as the "coders."

The coder training process and implementation were as follows. First, each coder used procedures outlined in the *Coder Training Manual and Reference Guide for Conducting Meta-Analysis* (Arthur & Bennett, 1994) to code a single article prior to any formal training. Next, the coders met to discuss problems encountered in using the guide and the coding sheet, and to make changes to the guide and/or the coding sheet as required. The coders were each assigned 5 articles to code. After coding these 5 articles, a second coder meeting was held and the degree of convergence between the two coders was assessed. Discrepancies and disagreements related to the coding of the 5 articles were resolved using a consensus discussion.

After this second meeting, the coders coded a common set of 20 articles which were used to assess the degree of interrater agreement. The level of agreement obtained for the primary meta-analysis variables is presented in Table 4. As these results indicate, the level of agreement was generally high with a mean overall agreement of 90.83% (SD = 3.37).

TABLE 4

Interrater Agreement for Major Study Variables

Study Variable	Agreement (%)
d	85
N	85
Evaluation Criteria	
Reaction	100
Learning	100
Behavior	90
Results	100
Training Method	90
Trainee Characteristics	85
Skill/Task Characteristic(s)	
Psychomotor	90
Cognitive	90
Interpersonal	80
Implementation Quality Total Score	95
Methodological Rigor Total Score	90
OVERALL	90.83

Calculating the Effect Size Statistic

In meta-analysis, cumulating the effects across studies requires that outcomes from all studies be converted to a common metric (Hunter & Schmidt, 1990). The present study used the effect size statistic (*d*) as the common metric. The effect size, or *d* statistic, provides a measure of the strength of a treatment or independent variable (e.g., different training methods). The effect size statistic, *d*, represents the observed difference between the experimental and the control group in standard deviation units (Cohen, 1990). A positive *d* value indicates that the experimental group performed better than the control group on the dependent variable. Conversely, a negative *d* value indicates that the control group performed better than the experimental group and a zero *d* value indicates no difference between the groups. As shown in Formula 1, the *d* statistic is calculated as the difference between the means of the experimental (M_E) and control groups (M_C) divided by a measure of the variation (Cohen, 1988; Glass, 1976; Glass et al., 1981; Hunter & Schmidt,1990).

The measure of variation used in the present study, S_W, was the pooled, within-group standard deviation (see Hunter & Schmidt, 1990, p. 271).

$$d = \frac{M_E - M_C}{S_W} \tag{1}$$

Although Formula 1 calls for the means of "experimental" and "control" groups, many training evaluation studies report other statistics (e.g., correlations, *t* statistics, or univariate two-group *F* statistics). For these studies, which constituted approximately 40% of the total data points, the appropriate conversion formulas were used to convert them to *d*s (Glass et al., 1981; Hunter & Schmidt, 1990; Wolf, 1986). For studies that reported actual means and standard deviations of the experimental (trained) and control (untrained) groups, *d* effect sizes were calculated directly using these statistics. These studies constituted 60% of the total data points.

Analyses

Cumulating/Aggregating Effect Sizes Across Studies. Using Arthur, Bennett, and Huffcutt's (1995a; Huffcutt, Arthur, & Bennett, 1993) SAS PROC MEANS meta-analysis program, a mean sample-weighted effect size (*d*) was calculated for each level of the independent variables using Formula 2 below:

$$d = \frac{\sum d_i N_i}{N_T} \tag{2}$$

where *d* is the mean effect size; d_i is the effect size for each study; N_i is the sample size for each study; and N_T is the total sample size across all studies. Sample weighting assigns studies with larger sample sizes more weight and reduces the effect of sampling error since sampling error generally decreases as the sample size increases (Hunter & Schmidt, 1990).

As previously indicated, the *d* statistic is a standard deviation metric used to express the difference between treatment and control groups, typically in experimental studies. There may be instances where the sample sizes are very uneven due to subject attrition or other factors. In these situations, Hunter and Schmidt (1990) recommend "correcting" the mean *d* ($\bar{d}$) for the attenuating effect of unequal sample sizes. This is accomplished using a bias multiplier, denoted as "A", which is calculated as (Hunter & Schmidt, 1990, p. 281-283; 289):

$$A = 1 + (.75 / \bar{N} - 3)) \tag{3}$$

It should be noted that for sample sizes of 100 or larger, the bias multiplier will differ only trivially from 1.00. The corrected mean d ($\bar{d}$)and the standard deviation of the population effect sizes ($SD\delta$) are then obtained by dividing the mean d and standard deviation by the bias multiplier as presented in Formula 4 and Formula 5 below:

$$\delta = \bar{d} / A \tag{4}$$

$$SD\delta = Var(\delta)^{1/2} / A \tag{5}$$

where $Var(\delta)$ is the population variance.

Moderator Analysis. Next, for the assessment of each factor hypothesized to influence training effectiveness, studies were categorized into separate subsets according to the specified level of the factor. An overall, as well as a subset mean effect size, were calculated for each factor. In terms of meta-analysis, an influencing variable, or moderator, is defined as any variable that, by its inclusion in the analysis, accounts for, or adds to the explained variance of the analysis. This is somewhat different from the use of the term "moderator" in multiple regression. In multiple regression, a moderator is a variable that, while having a negligible correlation with a criterion, interacts with another variable in the prediction model so as to enhance the predictability of a criterion variable (Cohen & Cohen, 1983). Thus, in meta-analysis, an influencing (or moderator) variable or factor is identified if (a) the effect size variance is lower in the subsets than for the factor as a whole; and/or (b) the average effect size varies from subset to subset. In brief, if large differences are found between subsets of a given factor, then the factor can be considered to be a moderator variable.

In addition, the correlation approach to the search for moderator variables was used in several instances. In this approach, correlations between sample-weighted ds and the hypothesized moderator variables were computed and interpreted.

RESULTS

Overall Training Effectiveness

One of the primary goals of this study was to examine the effectiveness of organizational training. It was hypothesized (hypothesis 1) that there would be a positive overall effect size for training. However, this effect size was expected to be fairly small. The results from the overall assessment of training are presented in Table 5. To generate the overall, sample-weighted *d*, individual study *d*s for each of the 466 individual data points were aggregated. As presented in Table 5, the overall sample-weighted, corrected *d* (δ) across the 466 data points was .751. This result indicates that training is effective, although the magnitude of the *d* is not "fairly small" as hypothesized, but rather large (Cohen, 1992).

The standard deviation of the corrected overall *d* (*SD*δ = .540) is large enough to suggest the presence and operation of potential moderator variables (Hunter & Schmidt, 1990). Therefore, as hypothesized, there are likely to be a number of factors that potentially moderate the effectiveness of training.

TABLE 5

Meta-Analysis Results for Overall Training Effectiveness

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall TE	47605	466	.757	.751	.540	1.613	5.090	12.630	-.307	1.810

NOTE: TE = training effectiveness; mean *d* = sample-weighted *d*; δ = corrected mean *d*; *SD*δ = standard deviation of the corrected mean *d*. The confidence interval is used to assess the accuracy of the estimate of the mean effect size (δ). Specifically, it estimates the extent to which sampling error remains in the sample-weighted effect size.

Moderator Analyses

In order to examine the influence of potential moderators, separate meta-analyses were run for each subset of these variables. As previously discussed, a moderator variable or factor is identified if (a) the effect size variance is lower in the subsets than in the factor as a whole; and/or (b) the average effect-size varies from subset to subset. In other words, if large differences are found between subsets of a given factor and the overall effect-size, then that factor can be considered to be a moderator variable. Consequently, each subset of hypothesized moderator variables was examined.

One hypothesized potential moderator was the quality of the implementation process (hypothesis 2). This was based on the premise that a systematic needs assessment should provide a mechanism whereby the questions essential to successful training programs can be identified

and addressed. Assessing the training needs of the organization in terms of identifying job requirements to be trained, who needs the training, and the kind of training to be delivered should result in more effective training programs. Therefore, studies which reported any of the three needs assessment activities as part of the training development or implementation process (e.g., person, task, and organizational analysis) were seen as being of higher quality in terms of their implementation than those that did not report any needs assessment activities.

Implementation Quality. It was hypothesized that studies that were of higher implementation quality, as measured by reported needs assessment activities, would report larger effect sizes than those studies that did not. Table 6 presents the meta-analysis results for implementation quality. In order to assess the overall impact of *any* needs assessment activities, an overall sample-weighted, corrected *d* (δ) across the 32 studies that reported one or more needs assessment activities was computed.

TABLE 6

Meta-Analysis Results for Implementation Quality

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall TE	**47605**	**466**	**.757**	**.751**	**.540**	**-1.613**	**5.090**	**12.630**	**.307**	**1.810**
Overall IQ	**1292**	**32**	**.803**	**.787**	**.419**	**.050**	**3.560**	**38.164**	**.034**	**1.608**
IQ Level										
0	**46313**	**434**	**.755**	**.750**	**.543**	**-1.613**	**5.090**	**12.036**	**.314**	**1.815**
1	**965**	**19**	**.745**	**.733**	**.412**	**.050**	**3.560**	**33.356**	**.075**	**1.541**
2	**327**	**13**	**.974**	**.942**	**.385**	**.210**	**1.949**	**55.059**	**.188**	**1.696**

NOTE: The range of possible Implementation Quality (IQ) scores is 0 - 3. The "0" category is provided to show the number of studies that did not report any Implementation Quality information or activity. No study reported all 3 activities.

As presented in Table 6, the corrected *d* for studies reporting any needs assessment activities is larger (δ = .787) than for those studies that did not (δ = .750). In addition, the overall *d* for implementation quality is somewhat higher than the overall *d* reported for all training effectiveness studies. It is important to note that the studies reporting any needs assessment activities represented a fairly small percentage (6.87%) of all the studies in this meta-analysis. Table 6 also presents a breakdown of implementation quality scores and associated corrected *d*s for each level of IQ. As shown, there is a substantial increase in the corrected *d* for studies reporting two needs assessment activities over those reporting only one (δ = .942 and .733 respectively). In addition, the reduction in the standard deviations across the subsets (that is, for studies reporting one or two needs assessment activities; *SD*δ = .412 and .385 respectively)

supports the hypothesis that implementation quality is a moderator of training effectiveness. These results generally support the contention that conducting needs assessment activities will help improve the reported effectiveness of training. Specifically, the observed effectiveness of training appears to increase as the implementation quality also increases.

Evaluation Criteria. Hypothesis 3 postulated that as the criteria used for evaluating the training program became more distal (e.g., reactions versus results), the magnitude of the average, sample-weighted, corrected *d* (δ) would decrease accordingly. Specifically, it was hypothesized that training studies that used reaction measures would have larger overall *d*s than those that used other, more distal, criteria such as learning, behavior, and results. In addition, training studies that used learning criteria as measures of effectiveness, would have larger overall *d*s than those that used either behavior or results criteria. Furthermore, training studies that used behavioral criteria would have smaller overall *d*s than those that used either learning or reaction criteria. Further, training studies that used results criteria would report smaller effect sizes than those that used behavioral, learning, or reaction criteria. It was also hypothesized (hypothesis 5) that training studies which were found to be effective in terms of behavioral criteria would be found to be effective in terms of results criteria, although the number of studies reporting the use of both behavior and results criteria was expected to be small. Table 7 presents the meta-analysis results for the criteria used to evaluate training.

The results presented in Table 7 did not support hypothesis 3. The fully corrected *d*(δ) was expected to become smaller as the level of measurement moved from reactions to learning to behavior to results. However, as shown in Table 7, the magnitude of the corrected *d* is quite similar for reactions, learning, and behavior (δ = .636; δ = .649; and δ = .624 respectively). In addition, the magnitude of the corrected *d* for results criteria is substantially larger (δ = 1.207) than it is for any of the other "levels" of criteria. One explanation for the markedly higher corrected *d* for results criteria is that these criteria are typically measures of organizational effectiveness (e.g., market share, turnover, absenteeism, productivity, profitability). As such, the impact of a training program may, or may not, be related to change observed with results criteria. In other words, the magnitude of change in the results criteria may have little or nothing to do with the training intervention or changes in individual worker behaviors. Given the more macro level of measurement associated with results criteria, it is reasonable to expect that the magnitude of change in these criteria might be quite large, but possibly minimally related to a previous training program.

TABLE 7

Meta-Analysis Results for Training Effectiveness Levels of Criteria

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall TE	**47605**	**466**	**.757**	**.751**	**.540**	**1.613**	**5.090**	**12.630**	**.307**	**1.810**
"Level" of Criteria										
Reactions	**660**	**11**	**.644**	**.636**	**.311**	**.140**	**1.860**	**42.241**	**.026**	**1.250**
Learning	**21973**	**290**	**.656**	**.649**	**.634**	**-1.613**	**5.090**	**12.223**	**-.594**	**1.892**
Behavior	**15527**	**132**	**.628**	**.624**	**.287**	**-0.699**	**3.096**	**30.280**	**.061**	**1.187**
Results	**9445**	**33**	**1.211**	**1.207**	**.370**	**.070**	**2.130**	**10.824**	**.483**	**1.931**

For most of the levels of criteria shown in Table 7, there is a reduction in the *SD*δ (e.g., for reactions, behavior, and results). This reduction in the *SD*δ does support the assertion that level of criteria is a moderator of training effectiveness. However, the variation of the *d* from subset to subset does not appear to be systematic. That is, there does not appear to be a pattern of either increasing or decreasing variation as a function of criterion level.

To further explore hypothesis 3, each level of criteria was assigned a code related to the proximal or distal nature of the criteria with training. Thus, reaction criteria were assigned a code of "1", learning criteria were assigned a code of "2", behavior criteria were assigned a code of "3", and results criteria, as the most distal criteria for training effectiveness, were assigned a code of "4". A correlation of the observed *d* and its assigned code was computed. The resulting correlation, $r = -.1213$ (466), $p<.01$, although small, lends support to the proposed relationship between the proximal or distal nature of the criteria and the effectiveness of the training program. However, the correlational finding is somewhat contradictory to the findings presented in Table 7. One explanation for the negative correlation could be related to the actual variability associated with the minimum and maximum effect sizes observed for each level of criteria. In other words, although the overall, observed sample-weighted *d* for each level of criteria appear highly similar, the variability amongst the criteria within a level is greater than might be suggested by merely examining the overall observed *d* for each level of criteria.

Environmental Favorability. It was originally hypothesized (hypothesis 4) that training studies which accounted for the favorability of the post-training environment, in conjunction with behavioral criteria, would report higher effect sizes than studies that did not account for the post-

training environment. Although a limited number of studies discussed the impact of such things as supervisor support or opportunities to perform trained tasks, they did not provide sufficient information for appropriate coding. Therefore, this hypothesis could not be tested.

With respect to levels of criteria, it was also hypothesized (hypothesis 5), that studies that were found to be effective in terms of behavioral criteria would also be found to be effective in terms of results criteria. As hypothesized, the correlation of behavior and results criteria was positive, $r = .254$ (18), $p = .309$. The magnitude of the correlation is not surprising given the previous discussion of the nature of results criteria. Typically, behavioral criteria are related to measurement of some aspect of work performance. However, as previously discussed, the molar nature of typical results criteria reduces their sensitivity to change in actual work behaviors brought about by the training. To ensure that change at one level of criteria is manifest in observable change at subsequent levels, quantitative linkages of the criteria to one another must be established before the training begins. It is also important to note that the obtained correlation between behavior and results criteria is based on the small number of studies where the measurement of behavior *and* results was reported (k = 18). In fact, the number of studies actually reporting the measurement of any two or more levels of criteria was quite small (k = 20). This finding is consistent with other recent levels of criteria research (e.g., Alliger & Janak, 1989; Alliger, Tannenbaum, & Bennett, 1995).

Training Methods and Skill/Task Characteristics. Hypothesis 6 postulated that different training methods would be more or less effective as a function of the content to be trained. Certain training methods are more readily suited to providing training related to specific content than are others. It is important to match the content of the training with an appropriate method to achieve effective training. Similarly, the characteristics of the skills and tasks to be trained are also factors that could potentially influence the effectiveness of training. Prior to exploring the relationship of training methods and skill and task characteristics, the overall effectiveness of training by training method and by skill and task characteristics was examined. Table 8 presents the overall results for each training method. For convenience and ease of explanation, these techniques have been summarized into the two broad categories described in the literature review: (a) on-site methods, and (b) off-site methods (Wexley & Latham, 1991). Note that not all training methods discussed in the literature review are presented in Table 8; a given training method must have had at least four data points to be included.

With respect to training methods, the results presented in Table 8 indicate that the choice of training method is a potential moderator of the effectiveness of training.
This is evidenced by the fact that the effect size standard deviation is lower in the majority of the subsets than in for overall training effectiveness and the average effect-size varies substantially

from subset to subset. The corrected, sample-weighted *d*s range from 1.074 ($SD\delta$ = .425) for job aids to .391 ($SD\delta$ = .180) for equipment simulators. However, in most cases, the standard deviations are large enough to suggest that there might be additional moderators present.

TABLE 8

Meta-Analysis Results for Overall Training Method

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	$SD\delta$	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall										
TE	**47605**	**466**	**.757**	**.751**	**.540**	**-1.613**	**5.090**	**12.630**	**.307**	**1.810**
On-Site Methods										
Orient	**444**	**4**	**.401**	**.398**	**.265**	**.310**	**1.750**	**34.414**	**.122**	**.918**
Job Aids	**277**	**10**	**1.107**	**1.074**	**.425**	**-.324**	**1.900**	**48.450**	**.242**	**1.907**
Off-Site Methods										
Lecture	**31689**	**249**	**.855**	**.850**	**.504**	**-.699**	**3.911**	**11.933**	**.138**	**1.839**
A/V	**7691**	**126**	**.694**	**.686**	**.649**	**-.034**	**5.090**	**14.250**	**.587**	**1.960**
PI	**3578**	**40**	**.676**	**.670**	**.464**	**.036**	**4.024**	**18.110**	**.239**	**1.578**
Discuss	**16316**	**243**	**.512**	**.506**	**.583**	**-1.613**	**4.351**	**15.420**	**.637**	**1.649**
CAI	**4923**	**42**	**.444**	**.441**	**.413**	**-.324**	**2.077**	**17.076**	**.368**	**1.251**
Eq Sim	**1688**	**20**	**.394**	**.391**	**.180**	**.211**	**1.965**	**59.930**	**.038**	**.744**
Self-T	**4091**	**39**	**.450**	**.446**	**.421**	**-.025**	**2.278**	**18.113**	**.380**	**1.272**

NOTE: Orient = orientation; A/V = audio/visual; PI = programmed instruction; CAI = computer-assisted instruction; Discuss = Discussion; Eq Sim = equipment simulators; Self-T = self-taught.

The results presented in this table provide considerable information about the overall effectiveness of the various methods. Moreover, this information can be used to actually rank-order the methods in terms of their overall effectiveness. This rank-ordering capability is an important contribution as training developers have typically had to rely on subjective judgments of the potential effectiveness of different methods, in the absence of empirical data. As hypothesized, however, information related to the overall effectiveness of the different methods, does not provide any information about the specific skills and tasks for which a particular method is most appropriate.

Table 9 presents the results for each of the skill and task characteristics. The general classification scheme used in the present study for both skills and tasks includes psychomotor, cognitive, and interpersonal categories (Farina & Wheaton, 1973; Fleishman & Quaintance, 1984; Goldstein, 1993).

Cognitive skills and tasks are related to the thinking, idea generation, understanding, or knowledge requirements of the job. Training which has focused on these skills and tasks was

found to be the most effective (δ = .831). In addition, the corrected, sample-weighted *d* for cognitive skills and tasks was larger than the overall *d* for training. Psychomotor skills and tasks include the behavioral activities associated with a job. These skills and tasks are related to the hands-on or "doing" parts of the job. Training which has focused on these tasks was also found to be effective (δ = .639), although the number of data points was fairly low (k = 71).

TABLE 9

Overall Meta-Analysis Results By Skill/Task Characteristics

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall										
TE	**47605**	**466**	**.757**	**.751**	**.540**	**-1.613**	**5.090**	**12.630**	**.307**	**1.810**
Skill/Task Characteristic										
Cognitive	**34012**	**241**	**.835**	**.831**	**.519**	**-1.613**	**5.090**	**10.282**	**.187**	**1.849**
P-motor	**4278**	**71**	**.648**	**.639**	**.401**	**-.324**	**2.670**	**30.427**	**.147**	**1.426**
Inter-p	**9229**	**150**	**.513**	**.506**	**.586**	**-.699**	**4.351**	**16.470**	**.642**	**1.655**

NOTE: P-motor = psychomotor; Inter-p = interpersonal.

Finally, interpersonal skills and tasks are those which are related to interacting with others in a workgroup or with clients and customers. Training which focused on these skills and tasks was found to be the least effective (δ = .506), although this effect size is still reasonably large. However, there is sufficient variability in the corrected, sample-weighted *d* for cognitive, psychomotor, and interpersonal characteristics (e.g., *SD*δ =.519; *SD*δ = .401; and *SD*δ = .586, respectively) to suggest additional moderators as hypothesized.

Hypothesis 6 postulated that the match between training methods and skill and task characteristics would moderate the effectiveness of training. It is important to match the content of the training with an appropriate method to achieve effective training. Similarly, the characteristics of the skills and tasks to be trained are also factors that could potentially influence the effectiveness of training. It is probable that the nature of skills and tasks to be trained will influence the effectiveness of training. Table 10 provides results related to the effectiveness of different training methods for training different skill/task characteristics. Again, not every training method which was discussed in the literature review is shown in the table. This was due to the lack of training studies that utilized each method. The methods and skill/task characteristics identified in Table 10 are those for which there were at least four data points for that method and characteristic.

TABLE 10

Meta-Analysis Results for Training Method by Skill and Task Characteristics

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall TE	47605	466	.757	.751	.540	-1.613	5.090	12.630	.307	1.810
On-Site Methods										
Orient/ Cognitive	444	4	.401	.398	.265	.3101	.750	34.415	-.122	.918
Job Aid/ P-motor	160	8	.830	.795	.448	-.324	1.900	52.611	-.083	1.673
Off-Site Methods										
Overall Lecture	31689	249	.855	.850	.504	-.699	3.911	11.894	-.138	1.839
Lecture/ Cognitive	25901	124	.920	.917	.471	-.412	3.911	8.998	-.007	1.840
Lecture/ P-motor	2220	23	.483	.479	.281	.050	1.380	35.181	-.072	1.030
Lecture/ Inter-p	4352	99	.661	.649	.622	-.699	3.560	20.082	-.569	1.867
Overall A/V	7691	126	.694	.686	.649	.034	5.090	14.250	-.587	1.960
A/V/ Cognitive	3756	58	.758	.749	.724	.070	5.090	11.291	-.670	2.168
A/V/ P-motor	1834	30	.643	.634	.545	.050	2.670	18.923	-.434	1.703
A/V/ Inter-p	2086	78	.625	.617	.581	.034	3.560	18.181	-.523	1.756

TABLE 10 (Continued)

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall PI	3578	40	.676	.670	.464	.036	4.024	18.110	-.239	1.578
PI/ Cognitive	3213	31	.637	.633	.467	.140	4.024	15.765	-.282	1.547
PI/ Inter-p	365	9	1.011	.991	.253	.036	1.805	63.850	.496	1.486
Overall Discuss	16316	243	.512	.508	.583	-1.613	4.351	15.420	-.637	1.649
Discuss/ Cognitive	5894	92	.568	.561	.647	-1.613	3.911	13.516	-.707	1.830
Discuss/ P-motor	2603	35	.585	.579	.421	.140	2.670	24.172	-.246	1.404
Discuss/ Inter-p	7763	114	.434	.429	.560	-.699	4.351	16.170	-.669	1.528
Overall CAI	4923	42	.444	.441	.413	-.324	2.077	17.076	-.368	1.251
CAI/ Cognitive	4773	36	.438	.435	.411	-.025	2.077	15.493	-.371	1.241
CAI/ P-motor	150	6	.647	.626	.398	-.324	1.900	52.013	-.154	1.407
Overall Eq Sim	1688	20	.394	.391	.180	.211	1.965	59.930	.038	.744
Eq Sim/ Cognitive	402	11	.581	.568	0.000	.313	1.109	100.000	.568	.568
Eq Sim/ P-motor	1286	8	.336	.335	.195	.211	1.965	42.794	-.480	.717
Overall Self-T	4091	39	.450	.446	.421	-.025	2.278	18.113	-.380	1.272
Self-T/ Cognitive	3824	32	.439	.437	.438	-.025	2.278	15.208	-.422	1.295
Self-T/ Inter-p	267	7	.597	.585	0.000	.159	1.064	100.000	.585	.585

NOTE: Inter-p = interpersonal; P-motor = psychomotor; A/V = audio/visual; PI = programmed instruction; Discuss = discussion; CAI = computer-assisted instruction; Eq Sim = Equipment Simulators; Self-t = self-taught; Orient = orientation. Only methods for which for data points were found are included

The results presented in Table 10 illustrate that different training methods were found to be more effective for training certain skills and tasks than for others. Moreover, the variability in the corrected, sample-weighted *d*s and their respective standard deviations, indicates that the match between a given training method and the skills and tasks to be trained, moderates the effectiveness of training.

The presented results also demonstrate the differential impact of different training methods for different skills and tasks. As mentioned in the literature review, Carroll et al,. (1972) asked corporate training directors to rate the relative effectiveness of different training techniques for achieving certain training objectives.

As an example, their study found that lectures were rated as being among the least effective methods for training across all training objectives (Carroll et al., 1972). Lecture results presented in Table 10 would contradict this finding. In addition, the results for the lecture method support the contention that lectures are the most appropriate method for training cognitive skills and tasks (Wexley and Latham, 1991). In fact, the lecture appears to be a relatively effective training method for many of the skill and task characteristics.

As would be expected, the lecture was found to be most effective for cognitive skills and tasks ($\delta = .917$, $SD\delta = .471$). The lecture was also found to be effective for training interpersonal skills and tasks ($\delta = .850$, $SD\delta = .504$). This result is not surprising given that many types of focus groups and sensitivity training activities use the lecture as the primary training method. As might be expected, the lecture was found to be relatively less effective for training psychomotor skills and tasks ($\delta = .479$, $SD\delta = .281$). In cases where the content of the training is psychomotor in nature, there are other training methods which are more ideally suited to training this content.

In terms of on-site methods, only two methods, orientation training and job aids, were found to have sufficient data points to permit their inclusion in the table. According to Wexley and Latham (1991), on-site training methods are those used for providing information and skills to trainees at the work site. The training is conducted within the same physical environment as the actual work to be performed. Orientation training is used to identify and reinforce organizationally "appropriate" behaviors and norms necessary for advancement within the organization (Cascio, 1991). Job aids are materials which are routinely available to the employee in the work setting which help in the conduct of work. Wexley and Latham (1991) have proposed that the on-site methods are used to improve trainee self-awareness, job skills, and motivation. As such, these methods are useful for improving cognitive skills and for enhancing both psychomotor and interpersonal skills in workers.

On-site training method results presented in Table 10 illustrate fairly low effectiveness of orientation training for cognitive skills and tasks. The corrected, sample-weighted *d* for

orientation training and cognitive skills and tasks was δ = .398 (*SD*δ = .265), however, these results are based on only four studies. As evidenced by the small number of studies using this method, it has not been used as a training method very often. It is likely to have limited applicability for training a broad range of skills and tasks that can be more effectively trained using other methods summarized in the table.

Results for job aids indicate that they were found to be very effective for training psychomotor skills and tasks. This finding is very consistent with the purpose for job aids. They are typically used to provide training and to assist in the performance of job tasks. The corrected, sample-weighted *d* for job aids was δ= .795 (*SD*δ = .448).

Off-site methods offer a somewhat different approach to training. Training conducted using these methods is typically accomplished in an environment which is removed from the workplace (Wexley & Latham, 1991). Lectures and audiovisual techniques, for example, train employees by focusing primarily on the cognitive aspects of job skills. These methods, then, would be useful for improving cognitive skills or enhancing psychomotor skills in workers. For cognitive skills, Wexley and Latham (1991) proposed that lectures, audiovisual techniques, programmed instruction, teleconferencing, and corporate classrooms were the appropriate training methods.

Results presented in Table 10 provide some evidence for the appropriateness of these training methods for certain skills and tasks. As previously mentioned, the lecture was shown to be a method with a broad range of applicability for cognitive, interpersonal, and psychomotor skills and tasks. Further, audiovisual methods were also found to be an effective training method for all three skill and task characteristics. For cognitive skills and tasks, audiovisual methods were found to be highly effective (δ = .746, *SD*δ = .724). Audiovisual methods were also found to be effective for training psychomotor and interpersonal skills and tasks (δ = .634, *SD*δ = .545; and δ = .617, *SD*δ = .581 respectively).

As proposed by Wexley and Latham (1991), programmed instruction was found to be an effective training method for cognitive skills and tasks (δ = .633, *SD*δ = .467). However, programmed instruction was found to be highly effective for training interpersonal skills and tasks (δ = .991, *SD*δ = .253), although these results are based on only 9 data points. While this finding might be seen inconsistent with the purposes for programmed instruction, a number of studies highlighted and discussed by Wexley and Latham (1991) involved the use of this method for training supervisory management principles, such as interacting with employees and customer relations for sales personnel.

For psychomotor skills, Wexley and Latham (1991) proposed that computer-assisted instruction and equipment simulators were the most appropriate methods to use. The results

presented in Table 10 partially support this proposal. Computer-assisted instruction and equipment simulators were found to be somewhat effective methods for training psychomotor skills and tasks. Results for computer-assisted instruction and psychomotor skills and tasks were $\delta = .435$, $SD\delta = .398$. For equipment simulators, the results for psychomotor skills and tasks were somewhat lower than expected. The corrected, sample-weighted d (δ) was .335 and the $SD\delta$ was .195. Thus, although lower than expected, equipment simulators were found to be marginally effective for training psychomotor skills and tasks. Additionally, both methods were found to be effective for training cognitive skills and tasks. This is not particularly surprising for computer-assisted instruction, as it has been used for classroom training in theory and concepts in several studies. Equipment simulators typically have focused on psychomotor training for specific aspects of job performance. The observed δ for equipment simulators and cognitive skills and tasks was .568 ($SD\delta = 0.000$). Also, it is important to note that there were more studies where equipment simulators had been used for cognitive training than for psychomotor training.

These results might be explained by examining the nature of equipment simulators. While they typically have high physical fidelity with the actual performance task, they are assumed to have limited use for providing training on the cognitive aspects of the task. However, it may be the case that these techniques are used to form a fundamental link between behaviors and solutions for current problems, as well as provide an opportunity to explore strategic problem solving issues within a specific job performance context. As such, they could focus more on the combined effect of job behaviors and the cognitive processes that might lead to those behaviors.

Finally, two categories of training methods were added for the present study. These additional methods were discussion and self-teaching. Discussion methods are characterized by activities such as focused problem-solving groups or group interactions involving subject-matter experts facilitating a seminar on different aspects of work performance. Self-taught methods are those where the individual trainee is provided with materials for self-study as would be the case with a correspondence course where there is typically no formal instructor. Discussion methods were found to be fairly effective for all three categories of skills and tasks. For cognitive skills and tasks, the observed effect size (δ) for discussion methods was found to be .561 ($SD\delta = .647$); for psychomotor skills and tasks, the observed effect size (δ) was found to be .579 ($SD\delta = .421$); and for interpersonal skills tasks, the observed effect size (δ) was .428 ($SD\delta = .560$). In fact, the observed effect size (δ) for the discussion method was higher for cognitive and psychomotor task than it was for overall discussion. However, the $SD\delta$ actually increased for cognitive skills and decreased for psychomotor skills and tasks.

The self-taught method was found to be somewhat effective for training cognitive and interpersonal skills and tasks. For cognitive skills and tasks, the observed effect size (δ) was .437 ($SD\delta$ = .438) and for interpersonal skills and tasks, δ = .585 ($SD\delta$ = 0.000).

In summary, the results presented in Table 10 support hypothesis 6. The use of different training methods does influence the effectiveness of training. Certain training methods were found to be more ideally suited to providing training related to a specific content domain than are others. However, contrary to past research, the present results indicate that a wider variety of methods may, in fact, be applicable to a broad range of skill and task training. This is an important finding for training developers and planners. As part of the needs assessment process, they are faced with selecting appropriate methods to maximize effectiveness for different content. The process of selecting a training method has typically involved using subjective recommendations from instructional design experts. The results presented in Table 10 provide empirical evidence of the relative effectiveness of different methods for training skill and tasks characteristics. However, it is also important to point out that there is sufficient variability remaining within some of the training types to warrant additional moderator assessments of subcategories of those methods.

Trainee Characteristics. With respect to trainee characteristics, the present study attempted to use results from the meta-analysis to contribute to the resolution of the debate and the divergent views related to ATI. That is, are there systematic differences in the reported effect size for studies that considered trainee characteristics in training and those that did not. Table 11 presents the results for the analyses related to the subset of studies that reported accounting for trainee characteristics.

TABLE 11

Meta-Analysis Results for Trainee Characteristics

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall TE	**47605**	**466**	**.757**	**.751**	**.540**	**-1.613**	**5.090**	**12.630**	**.307**	**1.810**
No ATI	**45671**	**426**	**.764**	**.759**	**.542**	**-.463**	**5.090**	**12.035**	**-.304**	**1.821**
ATI	**1934**	**40**	**.584**	**.574**	**.457**	**-1.613**	**3.096**	**29.432**	**-.322**	**1.470**

NOTE: ATI = aptitude treatment interaction.

Based on the results presented in the table, studies that did not account for trainee characteristics had an overall, sample-weighted, corrected *d* of .759 ($SD\delta$ = .542), while those that did account for trainee characteristics obtained a *d* of .574 ($SD\delta$ = .457). It is important to note that although the observed *d* for studies that accounted for trainee differences is smaller than that for studies that did not, the *d* is based on a relatively small number of studies. Given that the overall, sample-weighted, corrected *d* for studies that accounted for trainee characteristics is smaller than the observed *d* for studies that did not, the importance of an overall impact of ATI on training effectiveness is questionable. However, the standard deviation is large enough in the studies that accounted for trainee differences to warrant an exploration of the impact of specific trainee characteristics on training effectiveness.

Research exploring general ability-related ATI have been found to be more consistent than those involving specific abilities (see Ghiselli, 1973; Lohman & Snow, 1984; Mumford, Weeks, Harding, and Fleishman, 1988; Snow & Yallow, 1982; Tyler, 1962). Such learner characteristics as aptitude, motivation, and reading grade level were found to be important determinants of training performance (measured by training outcomes) above and beyond the influence of training course content (Mumford et al., 1988).

Therefore, even though an overall impact related to ATI was not observed, the potential impact of more specific trainee characteristics which were addressed in the studies was examined. Studies that accounted for trainee characteristics were broken into subsets according to the specific characteristics that were considered in each study.

The results from these additional analyses are presented in presented in Table 12. As shown in the table, there is considerable variation in both the corrected, sample-weighted *d* and the standard deviation for a number of the specific characteristics, indicating that in certain instances, the specific characteristics might moderate the effectiveness of training. Although the sample-weighted corrected *d* for general aptitude (δ = .607, $SD\delta$ = .082) was larger than the observed overall *d* for studies that accounted for ATI, it was not larger than the overall *d* for studies that did not account for ATI. Further, an examination of the sample-weighted, corrected *d*s for age, gender, and experience characteristics, does not support the literature advocating the existence of specific ATI in training.

Additional analyses related to ATI, examined the methodological rigor of ATI studies as a means of explaining the mixed findings in the literature. As highlighted in the literature review, several explanations for the inconsistencies seen in the research have been proposed. ATI studies conducted prior to 1977 typically had 40 or fewer subjects per condition and, therefore, lacked sufficient statistical power to detect significant effects (Cohen, 1988; Schmidt & Hunter, 1978). The studies did not control for plausible alternative explanations of the

observed effects. For example, training time was allowed to vary along with achievement. Finally, ATI studies generally did not randomly assign subjects to conditions or use experimental or quasi-experimental designs.

TABLE 12

Meta-Analysis Results for Specific Trainee Characteristics

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall TE	**47605**	**466**	**.757**	**.751**	**.540**	**-1.613**	**5.090**	**12.630**	**.307**	**1.810**
ATI Overall	**1934**	**40**	**.584**	**.574**	**.457**	**-1.613**	**3.096**	**29.432**	**-.322**	**1.470**
General Aptitude	**886**	**20**	**.617**	**.607**	**.082**	**.140**	**1.343**	**93.554**	**.448**	**.766**
Experience	**159**	**6**	**.297**	**.287**	**.499**	**-.699**	**3.096**	**38.470**	**-.690**	**1.266**
Person	**192**	**5**	**.054**	**.053**	**1.022**	**1.613**	**1.490**	**9.180**	**-1.950**	**2.056**
Gender	**1151**	**15**	**.437**	**.433**	**.097**	**.014**	**.622**	**85.157**	**.243**	**.622**
Age	**447**	**11**	**.482**	**.473**	**.114**	**.196**	**1.109**	**88.831**	**.250**	**.685**

NOTE: ATI = aptitude treatment interaction; Person = personality. Only characteristics for which at least four data points were found are reported here. The total number of data points across all characteristic will sum to 57, thereby exceeding the 40 data points reported for ATI Overall. This is due to the fact that more than one trainee characteristic was examined. Therefore that data point is "counted" for each characteristic as appropriate.

Results from these supplemental analyses are provided in Table 13. From the results presented, there is no clear evidence that ATI studies that scored lower, in terms of their overall methodological rigor, were more effective than those that scored higher.

In fact, studies that scored lower were not found to be as effective as those that received the highest score of 6. Studies that scored lower, in terms of methodological rigor, did evidence larger standard deviations than those that scored higher. Finally, contrary to the positive-findings bias literature, ATI studies that were found to be the most rigorous were also found to evidence the largest observed effect-size ($\delta = 1.1145$, $SD\delta = .351$).

TABLE 13

Meta-Analysis Results for ATI Studies and Methodological Rigor

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
ATI										
Overall	**1934**	**40**	**.584**	**.574**	**.457**	**-1.613**	**3.096**	**29.432**	**-.322**	**1.470**
Methodological Rigor Score										
2	51	5	.932	.844	1.008	-.699	3.096	30.966	-1.131	2.819
3	264	6	.251	.246	.925	-1.633	1.490	9.772	-1.567	2.059
4	1180	25	.597	.589	0.000	.210	1.109	100.000	.589	.589
6	165	4	1.168	1.145	.351	.366	1.510	48.291	.458	1.833

NOTE: Only instances where there were at least four data points are reported in the table. Missing methodological rigor scores indicate less than four ATI studies for that score.

Study Design/Methodological Rigor. Similar to the exploration of the impact of ATI on training effectiveness, the present study sought to provide evidence to help resolve the debate over the impact of methodological rigor on the observed effectiveness of training. In other words, the reported effect size would be different for studies that were found to be less rigorous than for studies that were found to be more rigorous.

The results from analyses related to this exploration are presented in Table 14. In some cases, the observed effect size for a given methodological score is larger than that for overall training effectiveness. However, the standard deviations associated with these effect sizes are also very large. Given the considerable variability in both the corrected-sample-weighted *d* and their respective standard deviations, methodological rigor can be considered as a moderator of training effectiveness.

Although there is a relatively normal distribution of studies across the score values, there is no discernable pattern or systematic reduction in the observed *d* as a function of the methodological rigor score. The only notable exception to this assessment is for studies that scored "7". These studies evidence the smallest observed *d* (δ = .284) and standard deviation (*SD*δ = .088).

However, contrary to past results from positive-findings bias studies, there is no apparent, systematic reduction in the effectiveness of training as a function of increasing methodological rigor. These results are highly consistent with results obtained by Barrick and Alexander (1987) and might be also be due to the fact that most of the practitioner-oriented, or "popular-press" literature was excluded from the present study.

TABLE 14

Meta-Analysis Results for Methodological Rigor

TE Factor	Total sample size	Number of data points	Corrected Statistics Mean *d*	δ	*SD*δ	Min *d*	Max *d*	% Variance due to sampling error	95% confidence intervals Lower bound	Upper bound
Overall										
TE	**47605**	**466**	**.757**	**.751**	**.540**	**-1.613**	**5.090**	**12.630**	**.307**	**1.810**
Methodological Rigor Score										
1	106	4	.662	.642	.246	.215	1.661	72.796	.159	1.124
2	792	30	1.044	1.012	.907	-.699	4.351	17.570	-.766	2.790
3	11969	87	.677	.673	.338	-1.613	3.560	21.277	-.011	1.335
4	21201	179	.904	.898	.582	-.463	5.090	9.951	-.242	2.038
5	6959	95	.623	.617	.517	-.224	2.670	17.729	-.397	1.631
6	4833	57	.630	.624	.573	-.412	3.198	13.160	-.499	1.748
7	1652	13	.285	.284	.088	.054	.831	80.570	.112	.456

NOTE: The range of possible methodological rigor scores is 0 - 9. The minimum score obtained by any study was 1. Only 1 study received a score of 8 and is not reported in the table. No study reported all 9 criteria. The "overall" data for methodological rigor is the same as that for overall training effectiveness (TE).

To further explore the issue, a correlation between methodological score and the observed *d* was calculated. The correlation obtained was $r = -.103$ (466), $p < .05$. This correlation indicated that there was some relationship between methodological rigor and observed effectiveness and that the relationship was in the direction proponents of the positive-findings bias debate would expect. That is, as methodological rigor increases, the likelihood of observing an effect due to training is diminished. However, this result is considerably smaller in magnitude than results in other positive-findings bias research (see Guzzo, Jette, & Katzell, Kondrasuk, 1981; 1985; Terpstra, 1981).

Recently, Roberts and Robertson (1992) have proposed that evidence of positive-findings bias might be related to the actual criteria used to determine methodological rigor. They argued that there are sufficient differences in the criteria to warrant analysis at more micro levels (Roberts & Robertson, 1992). They proposed three alternative subsets of methodological criteria. Table 15 presents the composition of the overall methodological rigor measure and the three alternative measures proposed by Roberts and Robertson (1992): the six criteria measure, the sampling criteria measure, and the measurement criteria measure.

Roberts and Robertson (1992) found that the use of different combinations of methodological criteria produced different results, but were generally consistent with those of other recent studies (e.g., Barrick & Alexander, 1987; Bullock & Svyantek, 1983; Woodman & Wayne, 1985). In addition, Roberts and Robertson (1992) found only one instance where

positive-findings bias was evidenced when sampling criteria were used to evaluate methodological rigor.

TABLE 15

Composition of Methodological Rigor Measures

Measure	Criteria Included
Nine Criteria Measure	1. Sampling strategy reported 2. Representative sample size 3. Control group use 4. Random assignment to conditions 5. Pretest/Posttest design 6. Significance testing cutoff 7. Reliability ≥ .60 reported 8. Objective criteria used 9. Multivariate statistical analysis
Six Criteria Measure	1. Sampling strategy reported 5. Pretest/Posttest design 6. Significance testing cutoff 7. Reliability ≥ .60 reported 8. Objective criteria used 9. Multivariate statistical analysis
Sampling Criteria Measure	1. Sampling strategy reported 2. Representative sample size 3. Control group use 4. Random assignment to conditions
Measurement Criteria Measure	6. Significance testing cutoff 7. Reliability ≥ .60 reported 8. Objective criteria used 9. Multivariate statistical analysis

For the present exploration, correlations between each of the alternative measures and the observed *d* were calculated. For the six criteria measure, the resulting correlation with *d* was $r = .009$ (466), $p = .849$. The sampling criteria measure correlation was $r = -.165$ (419), $p < .05$. The measurement criteria correlation was $r = -.085$ (463), $p < .10$. These results support the findings from Roberts and Robertson (1992). That is, that positive findings was most evidenced when the sampling criteria measure was used. However, the majority of evidence presented in this section does not provide support for the existence of positive-findings bias.

DISCUSSION

The present study examined the influence of several factors upon training effectiveness. Meta-analytic procedures were applied to the extant training effectiveness literature to: (a) provide a quantitative "population" estimate of training effectiveness (across multiple primary studies); (b) determine the extent to which several hypothesized factors are moderators of training effectiveness; (c) provide a quantitative indicator of the relative impact of these factors as moderators; and (d) attempt to provide empirical evidence to help resolve debates over the existence and impact of aptitude-treatment interactions (ATI) and methodological rigor upon training effectiveness outcomes.

Although most training developers believe that training is effective, limited evidence of the cumulative impact of training has been amassed. The present study attempted to obtain evidence of the effectiveness of training and quantitatively cumulate the evidence to derive a "population" estimate of the overall effectiveness of training. The overall estimate was expected to be positive, albeit small in magnitude. This was not the case. The overall, corrected, sample-weighted effect size for training was found to be quite large, (δ = .751). As expected, however, the standard deviation of this "population" estimate was large enough to suggest the operation of a number of moderators.

Most of the study hypotheses related to the moderators were supported. One rather striking result was related to the use of training needs assessment in the training development process. This was defined as implementation quality for the present study. Anecdotal information would suggest that it is prudent to conduct a needs assessment prior to training, but most organizations do not recognize the importance of this step in the development process. This was evidenced by the extremely low percentage of studies that reported any needs assessment activities prior to training implementation (7%). The present study provided compelling evidence that the conduct of needs assessment prior to training has a substantial impact on training outcomes. This evidence will be valuable to human resource practitioners, consultants, and training developers who must convince their clients and management that needs assessment is a critical part of the training process.

In terms of criteria used to evaluate training, it was expected that training effectiveness would systematically vary as a function of the criteria chosen to measure effectiveness. Similar to past reviews of the training literature (e.g., Alliger & Janak, 1989; Alliger, Tannenbaum, & Bennett, 1995) there were a limited number of studies that reported using reaction measures (k = 11). As stated earlier, one explanation for this finding is that studies of training effectiveness that only report the use of reaction criteria are not likely to be published in the

empirical literature. Given the problems associated with the use of reaction measures as criteria, this finding is not surprising.

The vast majority of studies reported using either learning (k = 290) or behavior (k = 132) criteria. In addition, the present study found 33 studies that reported using results criteria. The data, however, did not support hypothesis 3. Contrary to what was expected, the overall observed *d* for each type/level of criteria was quite similar (δ = .636; δ = .649; δ = .624; δ= 1.207, respectively). This finding is inconsistent with the contention that as a more proximal measure of effectiveness, reactions should produce the largest observed effect size. Further, it has been suggested that the more distal the measurement of outcomes from the actual training program, the less likely that effectiveness outcomes will be evidenced (see Alliger & Janak, 1989). The present results related to criteria suggest that this may not be the case.

The relationship between behavior and results criteria was also explored in this study. Specifically, it was expected that training that was found to be effective in terms of behavior criteria would also be found to be effective for results criteria. However, an assessment of this assertion requires that a study report measuring both behavior and results criteria. Only 18 studies which reported the measurement of both criteria were found. Correlational analyses supported this hypothesis (r = .2543 (18), p = .309).

Although the results for the magnitude decrease of the criteria were not as expected, training developers must ensure that multiple criteria are used in any evaluation of training. It is crucial to be able to systematically link changes in workplace behaviors to the training program and to subsequent results measures. As discussed earlier, results measures are more macro than either learning or behavior measures. As such, they tend to measure a variety of organizationally relevant constructs that may have little or no relevance to the training program. Therefore, future evaluation studies should more adequately demonstrate the linkage amongst the criteria. Establishing these linkages should be accomplished as part of the needs assessment process prior to training development and implementation. The most appropriate criteria should be identified so that change resulting from the training can be demonstrated in subsequent job behaviors and in more macro organizational outcomes.

Environmental favorability of the transfer environment was expected to be a factor in the assessment of training effectiveness in terms of behavior and results criteria. As noted earlier, although some researchers have begun to explore the impact of environmental favorability upon training outcomes, there were no *published* studies addressing environmental favorability that were codeable for this meta-analysis. While the lack of studies addressing this factor may limit the conclusions of the present study, the impact of the post-training environment upon training outcomes must be addressed in future research.

The results for the effectiveness of different training methods for training difference skills and tasks were quite interesting. Contrary to what might have been expected, a much wider variety of training methods were found to be effective for different skills and tasks. In many cases, the training developer must make decisions regarding the training method to be used in a given situation. Results from this study would suggest that a much broader range of methods, with demonstrated effectiveness, is available for consideration. In addition, results from the analysis of the overall effectiveness of the various methods (see Table 8) can be used in a comparative assessment of the effectiveness of each method.

For example, a training developer has been asked to propose a strategic plan for future organizational training activities. The developer must identify the off-site training methods that would be the "best" overall candidates for training, since the exact nature of the training is not specifically known. Based on the results from the present study, the developer will be able to rank-order recommended training methods for consideration. In this example, the lecture would be the obvious choice since the observed *d* (δ) was .850 - a large effect size - and one that was larger than the observed effect size for overall training (δ = .751). The lecture would be followed by audio-visual methods (δ = .686), programmed instruction (δ = .670), and discussion (δ = .506). Further, the developer could also recommend the use of job aids as a method for any on-site training that might be required (δ = 1.074).

The rank-ordered methods could then be examined in terms of the costs associated with implementation and maintenance to further refine the recommendations. In addition, the training developer does not have to rely on subjective judgments about the "potential" effectiveness of each method (e.g., Carrol, Paine, & Ivancevich, 1972), but can systematically evaluate the various methods using the meta-analytic results.

Results from this study demonstrated that training for cognitive skills and tasks was found to be the most effective (δ = .831), and was found to be more effective than overall training. In addition, training for psychomotor skills and tasks was also found to be effective (δ = .639). Finally, training for interpersonal skills and tasks was found to be somewhat less effective (δ = .506). As was expected, the standard deviations for each skill and task characteristic were large enough to suggest the presence of additional moderators.

The strategic training planning example can be extended to the discussion of training different skills and tasks. Specifically, once the rank-ordered list of effective methods has been developed, it might be further refined by examining results from this study related to the match between training method and skill and task characteristics.

Similar to the previous discussion of the effectiveness of different training methods, the relative effectiveness of training for different skills and tasks does not provide a complete picture

of the match between training methods and skill and tasks characteristics. Moreover, based on the strategic planning example, it would be more beneficial for training planning purposes to be able to match training methods to the content for which they would be most appropriate.

Therefore, using meta-analytic procedures, it was possible to determine the relative effectiveness of different training methods for different skills and tasks. That is, for a given skill and task content, a quantitative indicator of which method would be the most effective, was derived using meta-analytic procedures. Table 16 summarizes these results. These results indicate that the lecture would be the training method of choice for training cognitive skills and tasks. In addition, it would also be an effective method for training interpersonal skills and tasks. The lecture would be least effective for training psychomotor skills and tasks. The most effective method for training psychomotor skills and tasks appears to be on-site job-aids. Finally, the most effective method for training interpersonal skills and tasks was programmed instruction. Although this might seem counterintuitive, Wexley and Latham (1991) point out that this method has been used extensively to train such interpersonal content as employee supervisory skills and customer relations.

The summary of the results for training method and skill and task characteristics presented in Table 16 provides quantitative information which can be used by the practitioner to determine which training method would be most appropriate for training certain skills and tasks. Moreover, training method selection decisions can be based on empirical findings instead of subjective estimates of effectiveness. These results are also useful for the training researcher. Specifically, even though the effectiveness of the different methods for a variety of skills and tasks has been demonstrated, there is sufficient evidence from the magnitude of the standard deviations associated with the effect sizes to warrant additional investigation.

While evidence of the overall effectiveness of the methods was demonstrated in this study, the results do not provide information on exactly why that method would be more effective for some skills and tasks as opposed to others. Future research should attempt to identify what instructional attributes of a method impact the effectiveness of that method for different training content. Finally, the skill and task categories used in this study were quite broad. Future research might explore the specific nature of the skills or tasks that facilitate the effective use of more than one method. In addition, studies examining the differential effectiveness of various training methods for the same content and a single training method across a variety of skills and tasks should be conducted.

TABLE 16

Relative Effectiveness of Different Training Methods for Skill and Task Characteristics

Skill/Task Category	Training Method: Rank-Ordered Effectiveness for Each Skill/Task Characteristic	δ	*SD*δ
Cognitive	Lecture	.917	.471
	Audio Visual	.749	.724
	Programmed Instruction	.633	.467
	Equipment Simulator	.568	.000
	Discussion	.561	.647
	Self-Taught	.437	.438
	CAI	.435	.411
	Orientation	.398	.265
Psychomotor	Job Aid	.795	.448
	Audio Visual	.634	.545
	CAI	.626	.398
	Discussion	.579	.421
	Lecture	.479	.281
	Equipment Simulator	.335	.195
Interpersonal	Programmed Instruction	.991	.253
	Lecture	.649	.622
	Audio Visual	.617	.581
	Self -Taught	.585	.000
	Discussion	.429	.560

This study also attempted to provide empirical evidence to help resolve the debate over the existence of ATI. Although the literature related to ATI has been somewhat mixed, the results presented in this study do not support the existence of ATI. That is, based on the findings presented, trainee characteristics were not found to be a critical factor which influenced the effectiveness of training. An assessment of the potential pool of trainees should still be undertaken as part of the needs assessment process. This should be a routine part of any training implementation. However, it is probable that in most instances, the actual training intervention will simply "swamp" or overwhelm the influence of trainee characteristic effects. Further results from this analysis indicated that most of the studies which reported addressing trainees characteristics were found to be methodologically well-designed.

Finally, results from the present study did not provide evidence of positive-findings bias in the training effectiveness literature. In addition, the majority of the training studies included in the present study were of reasonable methodological quality. That is, most of the studies (74%) received a methodological score of at least "4". In addition, there appeared to be no systematic reduction in the observed effect size as a function of the score. The only exception to this finding

was for studies that received a score of "7". Although subsequent correlation analyses revealed a small, negative relationship between effectiveness and methodological rigor, the majority of the empirical evidence obtain in this study does not support the existence of positive-findings bias.

Also, as Roberts and Robertson (1992) found, there was a significant negative correlation between their sampling criteria measure and effectiveness. This finding supports their contention that evidence of positive-findings bias is related to the criteria chose for assessing methodological rigor. Taken together, results from the present study do not support the existence of positive-findings bias. This is especially significant as this is the first instance where a different content domain was used for the assessment of positive-findings bias. Past research related to positive-findings bias has focussed strictly on the organizational development (OD) literature (see Barrick & Alexander, 1987; Bullock & Svyantek, 1983; Woodman & Wayne, 1985). Results from the present study provide empirical evidence from a different content domain and suggest that positive-findings bias probably does not exist or, at the very least, does not generalize beyond the OD literature. The present results replicate findings from Roberts and Robertson (1992) and also provide information which is useful to the resolution of the debate related to positive-findings bias.

These results should not be taken as a justification for practitioners to routinely design poor training effectiveness studies. Results obtained from poorly designed studies should never be interpreted as evidence of an effective training program when, in fact, these findings are potentially spurious and based upon method bias. Every attempt should be made to design good training and to evaluate it appropriately. In fact, the criteria used to evaluate the extant literature in terms of methodological rigor serves as an excellent checklist of considerations for the design and evaluation of a training program.

Study Limitations

There were a number of limitations in the present study. First, one of the major factors hypothesized to influence the effectiveness of training was not examined. That factor, environmental favorability could not be examined due to the absence of codeable studies that accounted for, or addressed favorability in assessing the effectiveness of training. This is a limitation in the extant published literature that is related to the relatively recent recognition that the post-training environment might play a key role in the transfer of trained information beyond the training situation. In addition, researchers have recently begun to identify and attempt to measure key characteristics of the post-training environment that might impact training. It is anticipated that codeable, empirical studies examining the favorability of the post-training environment and its impact on training outcomes will be forthcoming.

Second, there were a number of instances where the effectiveness of specific training methods could not be assessed. This was due to an absence of sufficient empirical studies (data-points) for inclusion and analysis. In some cases, this absence was due to the fact that the specific method was not described in the published studies with sufficient detail to permit its classification as that method. In other cases, the method identified was a relatively new method with few published effectiveness studies or a method which was never widely applied.

Third, this study focused on fairly broad factors and their impact on training effectiveness. Although a number of levels within these factors were identified, a priori, and examined in the study, in most cases of the moderator analyses, the standard deviation of δ was large enough to suggest the existence of other moderators. In anticipation of this occurrence, specific and descriptive information about each of the factors was coded. Future research might examine and further categorize these descriptions for additional analyses. In the present study however, the choice of factors/moderators and levels within the moderators was theoretically and/or conceptually driven. Therefore, post hoc explorations of additional moderators was not attempted.

Fourth, this study focused primarily on quantifying the impact of each of the hypothesized factors meta-analytically. While this approach has provided considerable information related to the impact of the factors on the observed effectiveness of training, the question of the *relative* impact of each factor might warrant future investigation. In regression terms, one might be interested in the extent to which each factor contributes variance to the overall prediction of training effectiveness. For the purposes of the present study, this analysis was not accomplished.

SUMMARY AND CONCLUSIONS

Summary

Overall, the results from this study are especially important for empirically summarizing the current state of the training effectiveness literature. The majority of the hypotheses proposed in this study were supported.

First, training was found to be more effective than expected. This study obtained a "population" estimate of the effectiveness of training ($\delta = .751$). In addition, there was sufficient evidence ($SD\delta = .540$) to support a search for hypothesized moderators of training effectiveness. These hypothesized moderators included the implementation quality of the program, the criteria used to evaluate the training, the methods used for training, and the skill and task characteristics to be trained. In addition, it was hoped that results from this study might help to resolve issues associated with the impact of aptitude treatment interactions and the existence of positive-findings bias associated with the methodological rigor or quality of the evaluation study.

Second, implementation quality, as defined by the reported conduct of a needs analysis, was found to be a significant moderator of training effectiveness. Although the majority of studies (93%) did not report any needs assessment activities, those that did report these activities were found to be markedly more effective than those that did not.

Third, results related to the criteria used to evaluate training were not as expected. The effect size for training did not systematically vary as a function of the "level" of criteria. In addition, the effect size for results criteria was substantially larger than expected. This was most likely due to the more macro nature of measurement typically associated with results criteria.

The proposed impact of posttraining environmental factors could not be explored in this study. This was due to the lack of codeable studies exploring these factors.

Fourth, different training methods were found to be effective for different skills and tasks, and therefore functioned as moderators of training effectiveness. In addition, the present study provided quantitative indicators of the relative overall effectiveness of a variety of training methods, the overall effectiveness of training for the various skills and tasks to be trained, and the relative effectiveness of the methods for training specific skills and tasks.

Finally, the present study found no empirical evidence to support the existence of aptitude treatment interactions or the presence of positive-findings bias in studies of training effectiveness.

Conclusions

Results from this study will be useful to training practitioners and to researchers. For practitioners, these results provide quantitative information related to the impact of several moderators of training effectiveness. By examining the results for the hypothesized factors, training planners and developers should be able to make better decisions regarding which method to use and when each method would be most appropriate to use should be. In addition, the evidence presented in this study should facilitate the justification of the approach to be taken in the development and evaluation process.

Findings from this study also identify several additional areas for training research. Although a number of factors were found to moderate the effectiveness of training, further explorations related to how they influence training effectiveness should be conducted. Finally, there were several instances of a small number of studies or no studies for a particular factor. These "non-findings" in terms of the available literature are, in fact, findings nonetheless and demonstrate fruitful areas for future research in training effectiveness.

REFERENCES

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*, 3-27.

Ackerman, P. L., Sternberg, R. J., & Glaser, R. (Eds.) (1989). *Learning and individual differences: Advances in theory and research.* New York: W. H. Freeman.

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology, 41*, 331-342.

Alliger, G. M., Tannenbaum, S. I., & Bennett, W., Jr. (1995, August). A meta-analysis of levels of criteria in training evaluation. In M. S. Teachout (Chair), *Meta-analytic investigations of training effectiveness.* Symposium conducted at the annual conference of the American Psychological Association, New York, NY.

Arthur, W., Jr., & Bennett, W., Jr. (1994). *Coder training manual and reference guide for meta-analysis.* Technical Report. Department of Psychology, Texas A&M University. College Station, TX.

Arthur, W., Jr., Bennett, W., Jr., & Huffcutt, A. I. (1994). Choice of software and programs in meta-analysis: Does it make a difference? *Educational and Psychological Measurement, 54*, 776-787.

Arthur, W., Jr., Bennett, W., Jr., & Huffcutt, A. I. (1995a). *Conducting meta-analysis using the PROC MEANS procedure in SAS: A reference manual and users' guide.* Manuscript submitted for publication. Texas A&M University. College Station, TX.

Arthur, W., Jr., Bennett, W., Jr., Stanush, P. L., & McNelly, T. (1995b, August). Skill retention and decay: A meta-analysis. In M. S. Teachout (Chair). *Meta-analytic investigations of training effectiveness.* Symposium conducted at the annual conference of the American Psychological Association, New York, NY.

Arthur, W., Jr., Young, B. S., Jordan, D., & Shebilske, W. L. (in press). Effectiveness of didactic and group training protocols: The influence of trainee interaction anxiety. *Human Factors.*

Arvey, R. D., & Cole D. A. (1989). Evaluating change due to training. In R. Katzell & Goldstein (Eds.), *Training and development in organizations* (pp. 25-62). San Francisco: Jossey-Bass.

Arvey, R. D., Cole, D. A., Hazucha, J. F., & Hartano, F. M. (1985). Statistical power of training evaluation designs. *Personnel Psychology, 38*, 493-508.

Austin, J. T., Humphreys, L. G., & Hulin, C. L. (1989). Another view of dynamic criteria: A critical reanalysis of Barrett, Caldwell, & Alexander. *Personnel Psychology, 42*, 583-597.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*, 63-105.

Bandura, A. (1986). *Social foundations of thought and action.* Englewood Cliffs, NJ: Prentice-Hall.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99*, 388-399.

Barclay, J. R., & DeMeers, S. T. (1982). Classroom climate, student characteristics, and achievement in secondary schools. *School Psychology Review, 11*, 370-376.

Barrett, G. V., Alexander, R. A., & Doverspike, D. (1992). The implications for personnel selection or apparent declines in predictive validities over time: A critique of Hulin, Henry, and Noon. *Personnel Psychology, 45*, 601-617.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology, 38*, 41-56.

Barrick, M. R., & Alexander, R. A. (1987). A review of quality circle efficacy and the existence of positive-findings bias. *Personnel Psychology, 40*, 579-592.

Berryman-Fink, C. (1985). Male and female managers' views of the communication skills and training needs of women in management. *Public Personnel Management, 14*, 307-313.

Bouffard-Bouchard, T. (1990). Influence of self-efficacy on performance in a cognitive task. *Journal of Social Psychology, 130*, 353-363.

Bracht, G. H. (1970). The relationship of treatment tasks, personological variables and dependent variables to aptitude-treatment interactions. *Review of Educational Research, 40*, 627-645.

Bullock, R. J., & Svyantek, D. J. (1983). Positive-findings bias in positive-findings bias research. *Academy of Management Proceedings, 43*, 221-224.

Burke, M. J., & Day, R. R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of Applied Psychology, 71*, 232-246.

Campbell, J. P. (1971). Personnel, training and development. *Annual Review of Psychology, 22*, 565-602.

Campbell, J. P. (1988). Training design for performance improvement. In J. P. Campbell, R. J. Campbell, & Associates (Eds.)., *Productivity in organizations* (pp. 177-216). San Francisco: Jossey-Bass.

Campbell, J. P., Dunnette, M. D., Lawler, E., & Weick, K. E. (1970). *Managerial behavior, performance, and effectiveness*. New York: McGraw Hill.

Carroll, S. J., Paine, F. T., & Ivancevich, J. J. (1972). The relative effectiveness of training methods - expert opinion and research. *Personnel Psychology, 25*, 495-510.

Cascio, W. F. (1982). *Costing human resources: The financial impact of behavior in organizations*. New York: Van Nostrand Reinhold Co.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Christal, R. (1974). *The United States Air Force occupational research project* (AFHRL-TR-73-75). Lackland Air Force Base, TX: Air Force Human Resources Laboratory.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: LEA.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: LEA.

Connell, D. B., Turner, R. R., & Mason, E. F. (1985). Summary of findings of the school health education evaluation: Health promotion effectiveness, implementation, and costs. *Journal of School Health, 55*, 316-321.

Corno, L., & Snow, R. E. (1985). Adapting teaching to individual differences among learners. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: MacMillan.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis*. New York: John Wiley & Sons.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040-1048.

Elliot, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*, 5-12.

Eurich, N. P. (1985). *Corporate classrooms*. Princeton, NJ: Carnegie Foundation.

Facteau, J. D., Dobbins, G. H., Russell, J. E. A., Ladd, R. T., & Kudisch, J. D. (1992). *Noe's model of training effectiveness: A structural equations analysis*. Paper presented at the 7th Annual Conference of the Society for Industrial and Organizational Psychology, Montreal, Canada.

Farina, A. J., Jr., & Wheaton, G. R. (1973). Development of a taxonomy of human performance: The task-characteristics approach to performance prediction. *JSAS Catalog of Selected Documents in Psychology, 3*, 26-27 (Ms No. 323).

Fleishman, E. A. (1955). Leadership climate, human relations training, and supervisory behavior. *Personnel Psychology, 6*, 205-222.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks*. Orlando, FL: Academic Press.

Ford, J. K., & Noe, R. A. (1987). Self-assessed training needs: The effects of attitudes toward training, managerial level, and function. *Personnel Psychology, 40*, 39-53.

Ford, J. K., Quinones, M., Sego, D. J., & Speer Sorra, J. S. (1992). Factors affecting the opportunity to perform trained tasks on the job. *Personnel Psychology, 45*, 511-527.

Gettinger, M., & White, M. A. (1979). Which is the stronger correlate of school learning, time to learn or measured intelligence? *Journal of Educational Psychology, 71*, 405-412.

Ghiselli, F. F. (1956). Dimensional problems of criteria. *Journal of Applied Psychology, 40*, 1-4.

Ghiselli, F. F. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*, 461-477.

Gist, M. E. (1989). The influence of training method on self-efficacy and idea generation among managers. *Personnel Psychology*, 42, 787-805.

Gist, M. E., Stevens, C. K., & Bavetta, A. G. (1991). Effects of self-efficacy and post-training intervention on the acquisition and maintenance of complex interpersonal skills. *Personnel Psychology, 44*, 837-861.

Glass, G. V. (1970). Open discussion of Dr. Messick's paper and Blommer's and Cohen's comments: Discussion (pp. 210-220). In M. C. Wittrock & D. C. Wiley (Eds.), *The evaluation of instruction.* New York: Holt, Rinehart, & Winston.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social science research.* Beverly Hills, CA: Sage.

Goldstein, I. L. (1980). Training in work organizations. *Annual Review of Psychology, 31*, 229-272.

Goldstein, I. L. (1993). *Training in organizations: Needs assessment, development, and evaluation* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Goldstein, I. L., & Bruxton, V. M. (1982). Training and human performance. In M. D. Dunnette & F. A. Fleishman (Eds.), *Human performance and productivity: Human capability assessment.* Hillsdale, NJ: Erlbaum.

Goldstein, I. L., & Musicante, G. R. (1986). The applicability of a training transfer model to issues concerning rater training. In E. A. Locke (Ed.). *Generalizing from laboratory to field settings* (pp. 83-98). Lexington, MA: Lexington Books.

Green, B. F., & Hall, J. A. (1984). Quantitative methods for literature reviews. *Annual Review of Psychology, 35*, 37-53.

Guzzo, R. A., Jackson, S. E., & Raymond, R. A. (1987). Meta-analysis analysis. *Research in Organizational Behavior, 9*, 407-442.

Guzzo, R. A., Jette, R. D., & Katzell, R. A. (1985). The effects of psychologically based interventions programs on worker productivity: A meta-analysis. *Personnel Psychology, 38*, 275-292.

Hand, H. H., Richards, M. D., & Slocum, J. W. (1973). Organizational climate and the effectiveness of a human relations program. *Academy of Management Journal, 16*, 185-195.

Hays, R. T., Jacobs, J. W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis. *Military Psychology, 4*, 63-74.

Huber, V. L. (1985). Training and development: Not always the best medicine. Personnel, *62*, 12-15.

Huffcutt, A. I., & Arthur, W., Jr. (1995). Development of a new outlier statistic for meta-analytic data. *Journal of Applied Psychology, 80*, 327-334.

Humphreys, L. G., Henry, R. A., & Noon, S. L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. *Psychological Bulletin, 107*, 328-340.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

James, L. R. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin*, 80, 75-83.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition [Monograph]. *Journal of Applied Psychology, 74*, 657-690.

Kaplan, R. M., & Pascoe, G. C. (1977). Humorous lectures and humorous examples: Some effects upon comprehension and retention. *Journal of Educational Psychology, 69*, 61-65.

Katzell, R., & Goldstein, I. (Eds.) (1989). *Training and development in organizations*. San Francisco: Jossey-Bass.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of the American Society of Training and Development, 13*, 3-9.

Kirkpatrick, D. L. (1987). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resource development*. New York: McGraw-Hill.

Kondrasuk, J. N. (1981). Studies in MBO effectiveness. *Academy of Management Review, 6,* 419-430

Latham, G. P. (1988). Human resource training and development. *Annual Review of Psychology, 39,* 545-582.

Lohman, D. F., & Snow, R. E. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology,* 76, 347-376.

Mathieu, J. E., Tannenbaum, S. I., & Salas, E. (1992). Influences of individual and situational characteristics on measures of training effectiveness measures. *Academy of Management Journal, 35,* 828-847.

McGehee, W., & Thayer, P. W. (1961). *Training in business and industry.* New York: Wiley.

Mumford, M. D., Weeks, J. L., Harding F. D., & Fleishman, E. A. (1988). Relationship between student characteristics, course content, and training outcomes: An integrative modeling effort. *Journal of Applied Psychology, 73,* 443-456.

Noe, R. A. (1986). Trainee's attributes and attitudes: Neglected influences on training effectiveness. *Academy of Management Review, 11,* 736-749.

Noe, R. A., & Schmitt, N. M. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology, 39,* 497-523.

O'Connor, E. J., Peters, L. H., Pooyan, A., Weekly, J., Frank, B., & Erenkranz, B. (1984). Situational constraints effects on performance, affective reactions, and turnover: A field replication and extension. *Journal of Applied Psychology, 69,* 663-672.

Pentz, M. A., Trebnow, E. A., Hansen, W. B., MacKinnon, D. P., Dwyer, J. H., Johnson, C. A., Flay, B. R., Daniels, S., & Cormack, C. (1990). Effects of program implementation on adolescent drug use behavior. *Evaluation Review, 14,* 264-289.

Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. *Academy of Management Review, 5,* 391-397.

Roberts, D. R., & Robertson, P. J. (1992). Positive-findings bias, and measuring methodological rigor, in evaluations of organizational development. *Journal of Applied Psychology, 77,* 918-925.

Rouiller, J. Z., & Goldstein, I. L. (1991). *Determinants of the climate for transfer of training*. Paper presented at the Sixth Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Saari, L. M., Johnson, T. R., McLaughlin, S. D., & Zimmerle, D. M. (1988). A survey of management training and practices in U.S. companies. *Personnel Psychology, 41*, 731-743.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*, 1173-1181.

Schmidt, F. L., & Hunter, J. E. (1978). Moderator research and the law of small numbers. *Personnel Psychology, 31*, 215-232.

Severin, D. (1952). The predictability of various kinds of criteria. *Personnel Psychology, 5*, 93-104.

Shute, V. J. (1992). Aptitude-treatment interactions and cognitive skill diagnosis. In Regian, J. W., & Shute, V. J. (Eds.), *Cognitive approaches to automated instruction* (pp. 15-48). Hillsdale, NJ: Lawrence Erlbaum.

Shute, V. J., & Regian, J. W. (1991, November). *Adaptivity in intelligent tutoring systems: Costs and benefits*. Invited address to the Conference on Intelligent Computer-Aided Training, Houston, TX.

Snow, R. E. (1986). Individual differences and the design of educational programs. *American Psychologist, 41*, 1029-1039.

Snow, R. E., & Yallow, F. (1982). Education and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence*. New York: Cambridge University Press.

Steiner, D. D., Lane, I. M., Dobbins, G. H., Schnur, A., & McConnell, S. (1991). A review of meta-analyses in organizational and human resources management: An empirical assessment. *Educational and Psychological Measurement, 51*, 609-626.

Streker-Seeborg, I., Seeborg, M. C., & Zegeye, A. (1984). The impact of nontraditional training on the occupational attainment of women. *Journal of Human Resources, 3*, 452-471.

Tannenbaum, S. I., Mathieu, J. E., Salas, E., & Cannon-Bowers, J. A. (1991). Meeting trainees' expectations: The influence of training fulfillment on the development of commitment, self-efficacy, and motivation. *Journal of Applied Psychology, 76*, 759-769.

Tannenbaum S. I., & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology, 43*, 399-441.

Taylor, M. S., Locke, E. A., Lee, C., & Gist, M. E. (1984). Type A behavior and faculty research productivity: What are the mechanisms? *Organizational Behavior and Human Performance, 34*, 402-418.

Terborg, J. R., Howard, G. S., & Maxwell, S. E. (1980). Evaluating planned organizational change: A method for assessing alpha, beta, and gamma change. *Academy of Management Review, 5*, 109-121.

Terpstra, D. E. (1981). Relationship between methodological rigor and reported outcomes in organization development evaluation research. *Journal of Applied Psychology, 66*, 542-543.

Thayer, P. W., & Teachout, M. S. (1993). *A climate for transfer model.* Unpublished manuscript, Brooks Air Force Base, TX: Armstrong Laboratory Human Resources Directorate.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques.* New York: Wiley.

Tracey, J. B., Tannenbaum, S. I., & Kavanaugh, M. J. (1995). Applying trained skills on the job: The importance of the work environment. *Journal of Applied Psychology, 80*, 239-252.

Tyler, L. E. (1965). *The psychology of human differences.* Englewood Cliffs, NJ: Prentice Hall.

Wanous, J. P., Sullivan, S. H., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74*, 259-264.

Wesley, B. E., Krockover, G. H., & Hicks, C. R. (1985). Locus of control and acquisition of computer literacy. *Journal of Computer-Based Instruction, 12*, 12-16.

Wexley, K. N. (1984). Personnel training. *Annual Review of Psychology, 35*, 519-551.

Wexley, K. N., & Baldwin, T. T. (1986). Posttraining strategies for facilitating positive transfer: An empirical exploration. *Academy of Management Journal, 29*, 503-520.

Wexley, K. N., & Latham, G. P. (1991). *Developing and training human resources in organizations* (2nd ed.). New York: Harper Collins.

Wightman, D. C., & Sistrunk, F. (1987). Part-task strategies in simulated carrier landing final-approach training. *Human Factors*, *29*, 254-254.

Williams, T. C., Thayer, P. W., & Pond, S. B. (1991). *Test of a model of motivational influences on reactions to training and learning*. Paper presented at the Sixth Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Newbury Park, CA: Sage Publications.

Woodman, R. W., & Wayne, S. J. (1985). An investigation of positive response bias in evaluations of organizational development interventions. *Academy of Management Journal*, *28*, 889-913.

APPENDIX A

REFERENCE LIST OF ARTICLES INCLUDED IN THE META-ANALYSIS

Albanese, R. (1967, January). A case study of executive development: Measurement of perceptions in a national restaurant association program. *Training and Development Journal*, 28-34.

Allen, J. A., & Buffardi, L. C. (1986). Maintenance training simulator fidelity and individual differences in transfer of training. *Human Factors*, *28*, 497-509.

Alliger, G. M., & Horowitz, H. M. (1989, April). IBM takes the guessing out of the testing. *Training and Development Journal*, 69-73.

Anderson, R. C., Faust, G. W., & Roderick, M. C. (1968). Overprompting in programmed instruction. *Journal of Educational Psychology*, *59*, 88-93.

Aronoff, J. & Litwin, G. H. (1971). Achievement motivation training and executive advancement. *Journal of Applied Behavioral Sciences*, *7*, 215-229.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, *72*, 567-572.

Attwood, D. A., & Wiener, E. L. (1969). Automated instruction for vigilance training. *Journal of Applied Psychology*, *53*, 218-23.

Barbee, J. R., & Keil, E. C. (1973). Experimental techniques of job interview training for the disadvantaged: Videotape feedback, behavior modification, and microcounseling. *Journal of Applied Psychology*, *58*, 209-213.

Bare, C. E., & Mitchell, R. R. (1972). Experimental evaluation of sensitivity training. *Journal of Applied Behavioral Science*, *8*, 263-276.

Bartol, K. M., & Martin, D. C. (1987). Managerial motivation among MBA students: A longitudinal assessment. *Journal of Occupational Psychology*, *60*, 1-12.

Basadur, M., Graen, G. B., & Scandura, T. A. (1986). Training effects on attitudes toward divergent thinking among manufacturing engineers. *Journal of Applied Psychology*, *71*, 612-617.

Bass, B. M., (1962). Reactions to "Twelve Angry Men" as a measure of sensitivity training. *Journal of Applied Psychology, 46*, 120-124.

Bass, B. M., Cascio, W. F., McPherson, J. W., & Tragash, H. J. (1976). PROSPER-training and research for increasing management awareness of affirmative action in race relations. *Academy of Management Journal, 19*, 353-369.

Bazerman, M. H., & Neale, M. A. (1982). Improving negotiation effectiveness under final offer arbitration: The role of selection and training. *Journal of Applied Psychology, 67*, 543-548.

Beatty, R. W. (1973). Blacks as supervisors: A study of training, job performance, and employers' expectations. *Academy of Management Journal, 16*, 196-206.

Beck, M. A., & Roblee, K. (1982). Teacher effectiveness training: A technique for increasing student-teacher interaction. *College Student Journal, 16*, 131-133.

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62*, 64-69.

Bierman, R., Carkhuff, R. R., & Santilli, M. (1972). Efficacy of empathic communication training groups for inner city preschool teachers and family workers. *Journal of Applied Behavioral Science, 8*, 188-202.

Birkenbach, X. C. (1986). Self-report evaluations of training effectiveness: Measuring alpha, beta, and gamma change. *South African Journal of Psychology, 16*, 1-7.

Blake, R. R., & Mouton, J. S. (1966). Some effects of managerial grid seminar training on union & management attitude toward supervision. *Journal of Applied Behavioral Science*, 2, 387-409.

Bolman, L. (1970). Laboratory versus lecture in training executives. *Journal of Applied Behavioral Science, 6*, 323-335.

Bottger, P. C. & Yetton, P. W. (1987). Improving group performance by training in individual problem solving. *Journal of Applied Psychology, 72*, 651-657.

Bouchard, T. J. (1972). Training motivation, and personality as determinants of the effectiveness of brainstorming groups and individuals. *Journal of Applied Psychology, 56*, 324-331.

Brenner, A. M. (1971). Self-directed T-groups for elementary teachers: Impetus for innovation. *Journal of Applied Behavioral Science, 7*, 327-341.

Bruning, N. S. (1987). Effects of exercise, relaxation, and management skills training on physiological stress indicators: A field experiment. *Journal of Applied Psychology, 72*, 515-521.

Bretz, R. D., & Thompsett, R. E. (1992). Comparing traditional and integrative learning methods in organizational training programs. *Journal of Applied Psychology, 77*, 941-951.

Brown, E. M. (1968). Influence of training method and relationship on the halo effect. *Journal of Applied Psychology, 52*, 195-199.

Bunker, K. A., & Cohen, S. L. (1977). The rigors of training evaluation: A discussion and field demonstration. *Personnel Psychology, 30*, 525-541.

Bunker, D. R., & Knowles, E. S. (1967). Comparison of behavioral changes resulting from human relations training laboratories of different lengths. *Journal of Applied Behavioral Science, 3*, 505-521.

Burke, R. J. (1969, August). A plea for a systematic evaluation of training. *Training and Development Journal*, 24-29.

Burnaska, R. F. (1976). The effects of behavior modeling training upon managers' behaviors and employees' perceptions. *Personnel Psychologist, 29*, 329-335.

Campion, M. A., & Campion, J. E. (1987). Evaluation of an interviewee skills training program in a natural field experiment. *Personnel Psychology, 40*, 675-691.

Canino, C., & Cicchelli, T. (1988). Cognitive styles, computerized treatments on athematics: Achievement and reaction to treatments. *Journal of Educational Computing Research, 4*, 253-264.

Canter, R. R., Jr. (1951). A human relations training program. *Journal of Applied Psychology, 35*, 38-45.

Caplan, L. J., & Schooler, C. (1990). Problem solving by reference to rules or previous episodes: The effects of organized training, analogical models, and subsequent complexity of experience. *Memory & Cognition, 18*, 215-227.

Carron, T. J. (1966). Human relations training and attitude change: A vector analysis. *Personnel Psychology, 17*, 403-424.

Cavanagh, P., & Jones, C. (1968). An evaluation of the contribution of a program of self-instruction to management training. *Programmed Learning and Educational Technology, 5*, 294-301.

Chanow-Gruen, K. J., & Doyle, R. E. (1983). The counselor's consultative role with teachers: Using the TET model. *Humanistic Education and Development, 22*, 16-24.

Chentnik, C. G., & Weatherford, P. A. (1974). Teaching management by management exception. *Academy of Management Journal, 17*, 90-100.

Clinton, B. J., & Torrance, E. P. (1986). S.E.A.M: A training program for developing problem identification skills. *Journal of Creative Behavior, 20*, 77-80.

Cohen, D., Whitmyre, J. W., & Funk, W. H. (1960). Effect of group cohesiveness and training upon creative thinking. *Journal of Applied Psychology, 44*, 319-322.

Connor, D. V. (1968). Teaching engineering students by machine and text. *Programmed Learning and Educational Technology, 5*, 129-36.

Crawford, K. S., Thomas, E. D., & Fink, J. J. (1980). Pygmalion at sea: Improving the work effectiveness of low performers. *Journal of Applied Behavioral Science, 16*, 482-505.

Crimando, W. & Baker, R. (1984, September). Computer-assisted instruction in rehabilitation education. *Rehabilitation Counseling Bulletin*, 50-54.

Davis, B. L., & Mount, M. K. (1984). Effectiveness of performance appraisal training using computer assisted instruction and behavior modeling. *Personnel Psychology, 37*, 439-452.

Decker, P. J. (1980). Effects of symbolic coding and rehearsal in behavior-modeling training. *Journal of Applied Psychology, 65*, 627-634.

Decker, P. J. (1982). The enhancement of behavior modeling training of supervisory skills by the inclusion of retention processes. *Personnel Psychology, 35*, 323-332.

Dillon, P. C., Graham, W. K., & Aidells, A. L. (1972). Brainstorming on a "hot" problem: Effects of training and practice on individual and group performance. *Journal of Applied Psychology, 56*, 487-490.

DiVesta, F. J. (1954). Instruction-centered and student-centered approaches in teaching a human relations course. *Journal of Applied Psychology, 38*, 329-335.

Dossett, D. L., & Hulvershorn, P. (1983). Increasing technical training efficiency: Peer training via computer-assisted instruction. *Journal of Applied Psychology, 68*, 552-558.

Driscoll, J. M., Meyer, R. G., & Schanie, C. F. (1973). Training police in family crisis intervention. *Journal of Applied Behavioral Science, 9*, 62-82.

Dugan, B. (1988). Effects of assessor training on information use. *Journal of Applied Psychology, 73*, 743-748.

Earley, P. C. (1987). Intercultural training for managers: A comparison of documenting and interpersonal methods. *Academy of Management Journal, 30*, 685-698.

Eliason, A. (1972). A study of the effects of quantitative training. *Academy of Management Journal, 52*, 147-158.

Evertson, C. M. (1989). Improving elementary classroom management: A school-based training program for beginning the year. *Journal of Educational Research, 83*, 82-90.

Fiedler, F. E. (1972). Predicting the effects leadership training and experience from the contingency model. *Journal of Applied Psychology, 56*, 114-119.

Fiedler, F. E., & Mahar, L. (1979a). The effectiveness of contingency model training: A review of the validation of LEADER MATCH. *Personnel Psychology, 32*, 45-62.

Fiedler, F. E., & Mahar, L. (1979b). A field experiment validating contingency model leadership training. *Journal of Applied Psychology, 64*, 247-254.

Fisher, J. D., Silver, R. C., Chinsky, J. M., Goff, B., Klar, Y., & Zagieboylo, C. (1989). Physiological Effects of participation in a large group awareness training. *Journal of Consulting & Clinical Psychology, 57*, 747-755.

Fotheringhame, J. (1984). Transfer of training: A field investigation of youth training. *Journal of Occupational Psychology, 57*, 239-248.

Fotheringhame, J. (1986). Transfer of training: A field study of some training methods. *Journal of Occupational Psychology, 59*, 59-71.

Frayne, C. A., & Latham, G. P. (1987). Application of social learning theory to employee self-management of attendance. *Journal of Applied Psychology, 72*, 387-392.

French, J. R. P., Sherwood, J. J., & Bradford, D. L. (1966). Change in self-identity in a management training conference. *Journal of Applied Behavioral Science, 2*, 210-218.

Frew, D. R. (1987). Effects of exercise, relaxation, and management skills training on physiological stress indicators: A field experiment. *Journal of Applied Psychology, 72,* 515-521.

Friedlander, F. (1967). The impact of organizational training laboratories upon the effectiveness and interaction of ongoing work groups. *Personnel Psychology, 20,* 289-307.

Friedlander, F., & Greenberg, S. (1971). Effect of job attitudes, training, and organizational climate on performance of the hard-core unemployed. *Journal of Applied Psychology, 55,* 287-295.

Ganster, D. C., Williams, S., & Poppler, P. (1991). Does training in problem solving improve the quality of group decisions? *Journal of Applied Psychology, 76,* 479-483.

Gavales, D. (1966). Effects of combined counseling and vocational training on personal adjustment. *Journal of Applied Psychology, 50,* 18-21.

Gilbert, J., Campbell, H. G., & Oliver, A. E. (1963, May). An evaluation of interdepartmental training with objective tests. *Training Directors,* 46-55.

Gist, M. E. (1989). The influence of training method on self-efficacy and idea generation among managers. *Personnel Psychology, 42,* 787-805.

Gist, M. E., Bavetta, A. G., & Stevens, C. K. (1990). Transfer training method: Its influence on skill generalization, skill repetition, and performance level. *Personnel Psychology, 43,* 501-523.

Gist, M., Rosen, B., & Schwoerer, C. (1988). The influence of training method and trainee age on the acquisition of computer skills. *Personnel Psychology, 41,* 255-263.

Gist, M. E., Schwoerer, C., & Rosen, B. (1989). Effects of alternative training methods of self-efficacy and performance in computer software training. *Journal of Applied Psychology, 74,* 884-891.

Gliessman, D. H., Pugh, R. C., Brown, L. C., Archer, A. C., & Snyder, S. J. (1989). Applying a research-based model to teacher skill training. *Journal of Educational Research, 83,* 69-81.

Hahn, D. C., & Dipboye, R. L. (1988). Effects of training and information on the accuracy and reliability of job evaluations. *Journal of Applied Psychology, 73,* 146-153.

Hand, H. H., & Slocum, J. W., Jr. (1972). A longitudinal study of the effects of a human relations training program on managerial effectiveness. *Journal of Applied Psychology, 56*, 412-417.

Hand, H. H., & Slocum, J. W., Jr. (1970). Human relations training for middle management: A field experiment. *Academy of Management Journal, 13*, 403-410.

Hand, H. H., Richards, M. D., Slocum, J. W., Jr. (1973). Organizational climate and the effectiveness of a human relations training program. *Academy of Management Journal, 16*, 185-195.

Harris, E. F., & Fleishman, E. A. (1955). Human relations training and the stability of leadership patterns. *Journal of Applied Psychology, 39*, 20-25.

Harris, W. A., & Vincent, N. L. (1967). Comparison of performance of sales training graduates and non-graduates. *Journal of Applied Psychology, 51*, 436-441.

Harrison, J. K. (1992). Individual and combined effects of behavior modeling and the cultural assimilator in cross-cultural management training. *Journal of Applied Psychology, 77*, 952-962.

Hautaluoma, J. E., & Gavin, J. F. (1975). Effects of organizational diagnosis and intervention on blue-collar "blues". *Journal of Applied Behavioral Science, 11*, 475-496.

Hayes, W. G., & Williams, E. I. (1971, April). Supervisory training: An index of change. *Training and Development Journal*, 34-38.

Hedberg, R., Steffen, H., & Baxter, B. (1965). Insurance fundamentals-A programmed text versus conventional text. *Journal of Educational Research, 75*, 22-25.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73*, 68-73.

Higgins, E., Moracco, J., & Danford, D. (1981a). Effects of human relations training on education students. *Personnel Psychology, 18*, 165-172.

Higgins, E., Moracco, J., & Danford, D. (1981b). Effects of human relations training on education students. *Journal of Educational Research, 75*, 22-25.

Holloman, C. R., & Hendrick, H. W. (1972). Effect of sensitivity training on tolerance for dissonance. *Journal of Applied Behavioral Science, 8*, 174-187.

House, R. J., & Tosi, H. (1963). An experimental evaluation of a management training program. *Academy of Management Journal, 6*, 303-315.

Hogan, P. M., Hakel, M. D., & Decker, P. J. (1986). Effects of trainee-generated versus trainer-provided rule codes on generalization in behavior-modeling training. *Journal of Applied Psychology, 71*, 469-473.

Hirumi, A., & Bowers, D. R. (1991). Enhancing motivation and acquisition of coordinate concepts by using concept trees. *Journal of Educational Research, 84*, 273-279.

Ivancevich, J. M. (1979). Longitudinal study of the effect of rater training on psychometric error in ratings. *Journal of Applied Psychology, 6*, 502-508.

Ivancevich, J. M. (1974). A study of a cognitive training program: Trainer styles and group development. *Academy of Management Journal, 17*, 428-439.

Ivancevich, J. M., & McMahon, J. T. (1976). Group development, trainer style, and carry-over job satisfaction and performance. *Academy of Management Journal, 19*, 395-412.

Ivancevich, J. M., & Smith, S. V. (1981). Goal setting interview skills training: Simulated and on the job analyses. *Journal of Applied Psychology, 66*, 697-705.

Jaffee, C. L., & Friar, L. (1969, August). Use of simulation in training disadvantaged employees for secretarial positions. *Training and Development Journal*, 30-37.

Johnson, D., & White, C. B. (1980). Effects of training on computerized test performance in the elderly. *Journal of Applied Psychology, 65*, 357-358.

Jones, D. H. (1965). Training industrial executives in reading: A methodology study. *Journal of Applied Psychology, 49*, 202-204.

Jones, D. H., & Carron, T. J. (1965). Evaluation of a reading development program for scientists and engineers. *Personnel Psychology, 18,* 281-295.

Justis, R. T., Kedia, B. L., & Stephens, D. B. (1978). The effect of position power and perceived Task Competence on Trainer Effectiveness: A partial utilization of Fiedler's contingency model of leadership. *Personnel Psychology, 31*, 83-93.

Kabanoff, B., & Bettger, P. (1991). Effectiveness of creativity training and its relation to selected personality factors. *Journal of Organizational Behavior, 12*, 235-248.

Kaplan, R. E., Lombardo, M. M., & Mazique, M. S. (1985). A mirror for managers: Using simulation to develop management teams. *Journal of Applied Behavioral Science, 21*, 241-253.

Kassarjian, H. H. (1965). Social character and sensitivity training. *Journal of Applied Behavioral Science, 1*, 433-440.

Kidd, J. S. (1961). A comparison of two methods of training in a complex task by means of task simulation. *Journal of Applied Psychology, 43*, 165-169.

Kindler, H. S. (1979). The Influence of a meditation-relaxation technique on group problem-solving effectiveness. *Journal of Applied Behavioral Science, 15*, 527-533.

Klein, J. D., & Freitag, E. (1991). Effects of using an instructional game on motivation and performance. *Journal of Educational Research, 84*, 303-308.

Komaki, J., Heinzmann, A. T., & Lawson, L. (1980). Effect of training and feedback: Component analysis of a behavioral safety program. *Journal of Applied Psychology, 65*, 261-270.

Lansky, D. T., Reddy, W. B., & Lansky, L. M. (1978). External (legal) coercion and internal commitment: A case study of an affirmative action training program in municipal government. *Journal of Applied Behavioral Science, 14*, 27-42.

Latham, G. P., & Frayne, C. A. (1989). Self-management training for increasing job attendance: A follow-up and a replication. *Journal of Applied Psychology, 74*, 411-416.

Latham, G. P., & Kinne, S. B., III (1974). Improving job performance through training in goal setting. *Journal of Applied Psychology, 59*, 187-191.

Latham, G. P., & Saari, L. M. (1979). The application of social learning theory to training supervisors through behavior modeling. *Journal of Applied Psychology, 64*, 239-246.

Lawshe, C. H., Bolda, R. A., & Brune, R. L. (1959). Studies in management training evaluation II: The effects of exposures to role playing. *Journal of Applied Psychology, 43*, 287-292.

Lefkowitz, J. (1972). Evaluating of a supervisory training program for police sergeants. *Personnel Psychology, 25*, 95-106.

Leister, A., Borden, D., & Fiedler, F. E. (1977). Validation of contingency model leadership training: Leader match. *Academy of Management Journal, 20*, 464-470.

Lennung, S., & Ahlberg, A. (1975). The effects of laboratory training: A field experiment. *Journal of Applied Behavioral Science, 11*, 177-188.

Lewis, A. (1990). Defining and establishing relationships between essential and higher order teaching skills. *Journal of Educational Research, 84*, 5-12.

Maher, C. A. (1981). Training of managers in program planning and evaluation: Comparison of two approaches. *Journal of Organizational Behavior Management, 3*, 45-56.

Mahoney, T. A., Jerdee, T. H., & Korman, A. (1960). An experimental evaluation of management development. *Personnel Psychology, 13*, 81-98.

Maier, N. R. F., Hoffman, L. R., & Lansky, L. M. (1960). Human relations training as manifested in an interview situation. *Personnel Psychology, 13,* 11-30.

Manz, C. C., & Sims, H. P. (1986). Beyond limitation: Complex behavioral and affective linkages resulting from exposure to leadership training models. *Journal of Applied Psychology, 71*, 571-578.

Mathieu, J. E., Martineau, J. W., & Tannenbaum, S. I. (1993). Individual and situational influences on the development of self-efficacy: Implications for training effectiveness. *Personnel Psychology, 46*, 125-147.

Mathieu, J. E., Tannenbaum, S. I., & Salas, E. (1992). Influences of individual and situational characteristics on measures of training effectiveness. *Academy of Management Journal, 35*, 828-847.

Mayo, G. D., & Longo, A. A. (1966). Training time and programmed instruction. *Journal of Applied Psychology, 50*, 1-4.

McDonald, B. A., Larson, C. D., Dansereau, D. F., & Spurlin, J. E. (1985). Cooperative dyads: Impact on text learning and transfer. *Contemporary Educational Psychology, 10*, 369-377.

McGehee, W., & Gardner, J. E. (1955). Supervisory training and attitude change. *Personnel Psychology, 8*, 449-460.

Metcalf, K. K., & Cruickshank, D. R. (1991). Can teachers be trained to make clear presentations? *Journal of Educational Research, 85*, 107-116.

Miller, S. G. (1990). Effects of a municipal training program on employee behavior and attitude. *Public Personnel Management, 19*, 429-441.

Mindak, W. A., & Anderson, R. E. (1971, May). Can we quantify an act of faith? *Training and Development Journal*, 2-10.

Miner, J. B. (1960). The effect of a course in psychology on the attitudes of research and development supervisors. *Journal of Applied Psychology, 44*, 224-232.

Miner, J. B. (1965). Studies in management education. New York: Springer.

Moffie, D. J., Calhoon, R., & O'Brien, J. K. (1964). Evaluation of a management development program. *Personnel Psychology, 17*, 431-440.

Mollenkopf, W. G. (1969). Some results of three basic skills training programs in an industrial setting. *Journal of Applied Psychology, 53*, 343-347.

Moore, L. F. (1967, October). Business games versus cases as tools of learning. *Training and Development Journal, 10*, 13-23.

Moscow, D. (1971). T-group training in the Netherlands: An evaluation and cross-cultural comparison. *Journal of Applied Behavioral Science, 7*, 427-448.

Mosel, J. N., & Tsacnaris, H. J. (1954). Evaluating the supervisor training program. *Journal of Personnel Administration and Industrial Relations, 1*, 99-104.

Moses, J. L., & Ritchie, R. J. (1976). Supervisory relationships training: A behavioral evaluation of a behavior modeling program. *Personnel Psychology, 29*, 337-343.

Myers, G. E., Myers, M. T., Goldberg, A., & Welch, C. E. (1969). Effect of feed-back on interpersonal sensitivity in laboratory training groups. *Journal of Applied Behavioral Science, 5*, 175-86.

Nadler, E. B., & Fink, S. L. (1970). Impact of laboratory training on sociopolitical ideology. *Journal of Applied Behavioral Science, 6*, 79-92.

Noe, R. A. & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology, 39*, 497-523.

O'Donnell, A. M., Dansereau, D. F., Rocklin, T., Huthecker, V. I., Young, M. D., Hall, R. H., Skaggs, L. P., & Lambiotte, J. G. (1988). Promoting functional literacy through cooperative learning. *Journal of Reading Behavior, 20*, 339-355.

O'Driscoll, M. P. (1987). Attitudes to the job and the organization among new recruits: Influences of perceived job characteristics and organizational structure. *Applied Psychology: An International Review, 36*, 133-145.

Oshry, B. I. & Harrison, R. (1966). Transfer from here-and-now to there-and-then: Changes in organizational problem diagnosis stemming from t-group training. *Journal of Applied Behavioral Science*, *2*, 185-198.

Petersen, P. B. (1972, April). Leadership training. *Training and Development Journal*, 38-42.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, *69*, 581-588.

Rala, A. P. (1966). A study of the educational value of business games. *Journal of Business*, *39*, 339-352.

Rawls, J. R., Perry, O., & Timmons, E. O. (1966). A comparative study of conventional instruction and individual programmed instruction in the college classroom. *Journal of Applied Psychology*, *50*, 388-391.

Reber, R. A., & Wallen, J. A. (1984). The effects of training, goal setting, and knowledge of results on safe behavior. A component analysis. *Academy of Management Journal*, *27*, 544-560.

Reddy, W. B. (1972). Interpersonal compatibility and self-actualization in sensitivity training. *Journal of Applied Behavioral Science*, *8*, 237-240.

Rohrbaugh, M. (1975). Patterns and correlates of emotional arousal in laboratory training. *Journal of Applied Behavioral Science*, *11*, 220-240.

Russell, J. S., Wexley, K. N., & Hunter, J. E. (1984). Questioning the effectiveness of behavior modeling training in an industrial setting. *Personnel Psychology*, *37*, 465-481.

Schwartz, H. A., & Long, H. S. (1969). A study of remote industrial training via computer-assisted instruction. *Journal of Applied Psychology*, *51*, 11-16.

Sharan, S., & Hertz-Lazarowitz, R. (1982). Effects of an instructional change program on teachers' behavior, attitudes, and perceptions. *Journal of Applied Behavioral Science*, *18*, 185-201.

Siegel, A. I., Richlin, M., & Federman, P. (1960). A comparative study of "transfer through generalization" and "transfer through identical elements" in technical training. *Journal of Applied Psychology*, *44*, 27-30.

Slocum, J. W., Jr. (1968, September). Sensitivity and self-awareness changes: An empirical investigation. *Training and Development Journal*, *22*, 38-47.

Smith, R. E. (1989). Effects of coping skills training on generalized self-efficacy and locus of control. *Journal of Personality and Social Psychology, 56*, 228-233.

Smith, P. E. (1976). Management modeling training to improve morale and customer satisfaction. *Personnel Psychology, 29*, 351-359.

Soloman, L. N., Berzon, B., & Davis, D. P. (1970). A personal growth program for self-directed groups. *Journal of Applied Behavioral Science, 6*, 427-451.

Steele, F. I. (1968). Personality and the "laboratory style". *Journal of Applied Behavioral Science, 4*, 25-45.

Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of roster memory issues. *Journal of Applied Psychology, 77*, 501-510.

Swezey, R. N., Perez, R. S., & Allen, J. A. (1988). Effects of instructional delivery system and training parameter manipulations on electromechanical maintenance performance. *Human Factors, 30*, 751-762.

Swezey, R. W., Perez, R. S., & Allen, J. A. (1991). Effects of instructional strategy and motion presentation conditions on the acquisition and transfer of electromechanical troubleshooting skill. *Human Factors, 33*, 309-323.

Tannenbaum, S. I., Mathieu, J. E., Salas, E., & Cannon-Bowers, J. A. (1991). Meeting trainees' expectations: The influence of training fulfillment on the development of commitment, self-efficacy, and motivation. *Journal of Applied Psychology, 76*, 759-769.

Trollip, S. R. (1979). The evaluation of a complex computer-based flight procedures trainer. *Human Factors, 21*, 47-54.

Valiquet, M. I. (1968). Individual change in a management development program. *Journal of Applied Behavioral Science, 4*, 313-325.

Vernardos, M. G., & Harris, M. B. (1973). Job interview training with rehabilitation clients: A comparison of videotape and role-playing procedures. *Journal of Applied Psychology, 58*, 365-367.

Viteles, M. S. (1959). "Human relations" and the "humanities" in the education of business leaders: Evaluation of a program of humanistic studies for executives. *Personnel Psychology, 12*, 1-28.

Warmke, D. L., & Billings, R. S. (1979). Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. *Journal of Applied Psychology, 64*, 124-131.

Welsh, P., Antoinetti, J. A., & Thayer, P. W. (1965). An industry wide study of programmed instruction. *Journal of Applied Psychology, 49,* 61-73.

Wesley, B. E., Krockover, G. H., & Hicks, C. R. (1985). Locus of control and acquisition of computer literacy. *Journal of Computer-Based Instruction, 12*, 12-16.

West, R. L., & Crook, T. H. (1992). Video training of images for mature adults. *Applied Cognitive Psychology, 6*, 307-320.

Wexley, K. N., & Baldwin, T. T. (1986). Posttraining strategies for facilitating positive transfer: An empirical exploration. *Academy of Management Journal, 29*, 503-520.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. (1973). Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology, 57*, 233-236.

Wiener, E. L. (1963). Knowledge of results and signal rate in monitoring: A transfer of training approach. *Journal of Applied Psychology, 47*, 214-222.

Wolfe, J. & Moe, B. L. (1973). An experimental evaluation of a hospital supervisory training program. *Hospital Administration, 18*, 65-77.

Woodman, R. W., & Sherwood, J. J. (1980). Effects of team development intervention: A field experiment. *Journal of Applied Behavioral Science, 16*, 211-237.

Zacker, J., & Bard, M. (1973). Effects of conflict management training on police performance. *Journal of Applied Psychology, 58*, 202-208.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology, 67*, 752-758.

APPENDIX B

STUDY CODING SHEET

Coder ID #: _____

Article Code #: _____

Study #: _____

Publication Year: _____

Reference: _____
_____
_____
_____
_____

Type of Publication [Please Check]

___ 7. Journal Article
___ 6. Book/Book Chapter
___ 5. Conference Paper/Presentation
___ 4. Technical Report
___ 3. Dissertation
___ 2. Master's Thesis
___ 1. Unpublished/Submitted Manuscript

Type of Study (Setting) [Please Check]

___ 1. Basic/Lab
___ 2. Applied/Real World

Study Statistics:

1)	More than one group? (Circle one)	Yes	No
2)	If "**No**" in item 1, what is the total N? (N_{TOT}):	_____	
3a)	If "**Yes**" in item 1, what is the number of subjects in the experimental group? (N_E):	_____	
3b)	What are the Mean and Standard Deviation for the experimental group? (M_E & SD_E):	Mean _____	SD _____
4a)	If "**Yes**" in item 1, what is the number of subjects in the control group? (N_C):	_____	
4b)	What are the Mean and Standard deviation for the control group? (M_C & SD_C):	Mean _____	SD _____
5)	Standard deviation within groups (SD_W):		_____
6)	If the d was obtained using a conversion formula, what was the test statistic that was converted?		_____
7)	Calculated d statistic: _____		

Calculations Work Page

Implementation Quality Indicators:

Organizational analysis reported:	(Yes) = 1	(No) = 0
Task analysis reported:	(Yes) = 1	(No) = 0
Person analysis reported:	(Yes) = 1	(No) = 0

Total: _____

Study Characteristics:

Probabilistic sampling strategy reported:	(Yes) = 1	(No) = 0
Experimental and Control Group:	(Yes) = 1	(No) = 0
Random Assignment:	(Yes) = 1	(No) = 0
N Subjects each Group: $N \geq 30$	(Yes) = 1	(No) = 0
Pretest/Posttest:	(Yes) = 1	(No) = 0
Cut Off for Significance: $\alpha \leq .05$	(Yes) = 1	(No) = 0
Dependent Variable Reliability Coefficient $\geq .60$	(Yes) = 1	(No) = 0
Objective criteria used:	(Yes) = 1	(No) = 0
Multivariate Analysis:	(Yes) = 1	(No) = 0

Total: _____

Task or Skill Characteristics: (Place a "1" in the blank for the task or skill characteristic and describe the characteristic; Please place a "0" in those blanks that *DO NOT* apply)

Cognitive: ____

(e.g., changing trainee's thoughts and ideas)

Describe characteristic:

-

-

-

Psychomotor: ____

(e.g., focusing on changes to behaviors related to the job)

Describe characteristic:

-

-

-

Interpersonal: ____

(e.g., focusing on relating training to the workplace environment; familiarization training)

Describe characteristic:

-

-

-

Other: Describe: _____

_____

_____

_____

Training Methods: (Place a "1" in the blank for all that apply; "0" otherwise):

On Site Methods:	1/0
- Career Development	____
- Orientation	____
- Job Aid	____
- Apprenticeship	____
- Coaching	____
- Job Rotation	____
- Other	

Describe: _____

_____

_____

_____

Off Site Methods:	
- Lecture	____
- Audiovisual	____
- Programmed instruction	____
- Teleconferencing	____
- Discussion	____
- Computer-assisted	____
- Equipment simulators	____
- Other	

Describe: _____

_____

_____

_____

Trainee Characteristics:

Study accounted for ATI? (Yes) = 1 (No) = 0

If yes, code for type: (Place a "1" in the blank for all that apply; "0" otherwise):

Pretest/posttest? (Yes) = 1 (No) =0

Education ___
General Aptitude
Cognitive ability ___
Psychomotor ability ___

Other: _____

_____

_____

_____

Experience ___
Motivation ___
Self-efficacy ___
Personality ___
Locus-of-control ___
Gender ___
Ethnicity ___
Type: ___
Age ___
SES ___
Other: ___

Describe: _____

_____

_____

_____

Evaluation Criteria: (Place a "1" in the blank for each that applies; "0" otherwise)

MI = Measurement Interval (if applicable) **in Days**:

	Crit Used? (Y/N) 1/0	MI
Reactions: (e.g.,Attitude survey, reward for work, etc.)	___	___
List/Describe Measure(s) Used:		
-		
-		
-		
Learning: (e.g., End of course test, work sample, etc.)	___	___
List/Describe Measure(s) used:		
-		
-		
-		
Subjective Behavior: (e.g., supervisor/peer ratings, quality assurance check, etc.)	___	___
List/Describe Measure(s) used:		
-		
-		
-		
Objective Behavior: (e.g., # units repaired, # of items produced, etc)	___	___
List/Describe Measure(s) used:		
-		
-		
-		

Evaluation Criteria Used (continued): (Place a "1" in the blank for each that applies; "0" otherwise):

MI = Measurement Interval (if applicable) **in Days**:

	Crit Used? (Y/N) 1/0	MI
Subjective Results: (e.g., job satisfaction, organizational climate, etc.)	___	___
List/Describe Measure(s) used:		
-		
-		
-		
Objective Results: (e.g., employee records, absenteeism, tenure, promotion in job, etc.)	___	___
List/Describe Measure(s) used:		
-		
-		
-		

Environmental favorability indicators: (Place a "1" in the blank for all that apply; "0" otherwise):

Was an assessment of the post-training environment conducted? Yes No

If Yes, complete the following specific items: (Yes = 1 / No = 0)

Measure used to gather ratings: a) Likert-scale b) Other:
Describe: _____

Environmental favorability indicators (continued): (Place a "1" in the blank for all that apply; "0" otherwise):

	Rating by category (report values)		
	1/0	Mean	SD
Ratings of Task components			
- Availability of tools/equipment:	___	___	___
- Availability of supplies:	___	___	___
- Familiarity with task:	___	___	___
- Availability of monetary resources:	___	___	___
- Time constraints:	___	___	___
- Working conditions:	___	___	___
Column totals:			

- Others: Describe: _____

_____

_____

_____

	Rating by category (report values)		
	1/0	Mean	SD
Ratings of Social components			
- Top management support:	___	___	___
- Supervisory support:	___	___	___
- Peer support:	___	___	___
- Subordinate support:	___	___	___
- Opportunities to perform:	___	___	___
- Feedback:	___	___	___
- Reinforcement:	___	___	___
Column totals:			

- Others: Describe: _____

_____

_____

Rater Training Study (1/0): _____